Reevaluating the role of verbalisation of faces for composite production: Descriptions of offenders matter!

Charity Brown* (1) psccbr@leeds.ac.uk

Emma Portch (2) eportch@bournemouth.ac.uk

Laura Nelson (3) laura.dalby@lancashire.pnn.police.uk

Charlie D. Frowd (4) cfrowd1@uclan.ac.uk

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

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

(3) Lancashire Constabulary, HQ, Saunders Lane, Hutton, PR4 5SB, UK

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

* Corresponding author: Charity Brown, School of Psychology, University of Leeds, Leeds, LS2 9JT, UK; Tel: (0044) 113 3435748; Email: psccbr@leeds.ac.uk.

Abstract

Standard forensic practice necessitates that a witness describes an offender's face prior to constructing a visual likeness, a facial composite. However, describing a face can interfere with face recognition, although a delay between description and recognition theoretically should alleviate this issue. In Experiment 1, participants produced a free recall description either 3-4 hours or 2 days after intentionally or incidentally encoding a target face, and then constructed a composite using a modern 'feature' system immediately or after 30-minutes. Unexpectedly, correct naming of composites significantly *reduced* following the 30-minute delay between description and construction for targets encoded 2 days previously. In, Experiment 2, participants in these conditions gave descriptions that were better matched to their targets by independent judges, a result which suggests that the 30-minute delay actually impairs access to details of recalled descriptions that are valuable for composite effectiveness. Experiment 3 found the detrimental effect of description delay extended to composites constructed from a 'holistic' face production system. The results have real-world but counterintuitive implications for witnesses who construct a face one or two days after a crime: after having recalled the face to a practitioner, an appreciable delay (here, 30 minutes) should be avoided before starting face construction.

Public significance statement

The research indicates that for witnesses (who may also be victims) who are invited to construct a face of a target one or two days after a crime, there should not be an appreciable delay between describing the face to a practitioner and face construction. Inserting a delay (here, of 30 minutes) can lead to a witness constructing a composite that is less readily identified.

Running head: Face recall and composite construction

Keywords: Facial composite, witness, victim, facial descriptions, verbal overshadowing, retention interval, encoding instructions.

Describing an individual’s facial features can interfere with later identification of that same individual, an effect of ‘verbal overshadowing’. In their seminal studies, Schooler and Engstler-Schooler (1990) showed participants a video or photograph of a male ‘offender’, and asked them to describe his facial features from memory (or not, in a control condition) and attempt to identify him from a line-up of similar-looking faces. Participants providing a description were less successful at identifying the ‘offender’.

Some researchers have failed to replicate this finding, and it may be that inconsistencies in establishing the verbal overshadowing effect for faces in part arise due to its sensitivity to certain boundary conditions. Indeed, Meissner and Brigham (2001) in a meta-analysis of 29 effect size comparisons of verbal overshadowing identified conditions associated with an increased likelihood of obtaining the effect. These included a short delay (< 10 minutes) between providing a description and viewing the line-up test, and provision of more detailed descriptions. More recently, Alonga et al. (2014) conducted a replication of two of the original studies by Schooler and Engstler-Schooler (1990), collecting and pooling data across many independent laboratories. A reliable effect of verbal overshadowing was found, but effect sizes of the two experiments differed substantially in magnitude. This difference was attributed to changes in experimental parameters, as a larger effect (i.e., more interference to memory) was observed when participants described the face immediately before the line-up test compared with 20 minutes before (immediately after encoding).

One theoretical account proposes that descriptions, particularly detailed descriptions, promote recall of misleading or imprecise information (e.g., Finger & Pezdek, 1999; Meissner, Brigham, & Kelley, 2001). With a short delay between the description and line-up test, this newly formed, but imprecise, verbal memory has a greater propensity to interfere with access to the original visual memory of the face, potentially acting as a source of retroactive interference. With longer delays, salience of the memory for the description may subside (Finger & Pezdek, 1999). In an alternative account, detailed descriptions may direct attention to individual features of the face (i.e., a featural analysis, Wells & Hryciw, 1984) at the expense of a more holistic / configural analysis that, in general, tends to be more effective for face recognition (Diamond & Carey, 1986). With a shorter delay, there may be less of an opportunity for participants to revert to a processing strategy that would be more suited to face recognition (e.g., a processing shift account, Schooler, 2002).

When participating in an identity parade (or line-up) in the real world, the optimum boundary conditions previously identified for observing verbal overshadowing are less likely to arise. Typically, the parade will take place days or weeks (or even months) after a witness has provided a statement. Witness descriptions are also generally sparse, as reported by police officers (Brown, Lloyd-Jones, & Robinson, 2008) or evidenced in archival studies (van Koppen & Lochun, 1997). However, when the identity of an offender is unknown and there are few (or no) leads in an investigation, a witness may work alongside a police practitioner to produce a sketch or use a computer system to construct a visual likeness of the offender's face, a facial composite. In this situation, conditions likely to elicit verbal overshadowing may well apply.

Computer software systems that produce composites may either emphasise selection and melding of individual facial features into a 'whole' face likeness (a feature-based process; e.g., E-FIT, PRO-fit, FACES), or involve selection and 'breeding' of whole faces to 'evolve' a composite (a holistic-based process; EvoFIT, E-FIT-V or EFIT-6, ID, see Frowd, 2017 for a detailed review). However, it is standard police practice for a witness, prior to creating a composite, to recall a detailed description of the offender's face. Both types of composite system involve face recognition as part of feature/face selection. In addition, witnesses must recognise when the emerging face sufficiently matches their memory of the offender. Thus, it would seem reasonable to suggest that, given current forensic practice, the process of constructing a composite face may be prone to the effect of verbal overshadowing.

To date, only one published study appears to have investigated this possibility. Frowd and Fields (2011) adopted a procedure that mimics the forensic situation whereby a witness constructs a composite of an offender who is *unfamiliar* to him or her, and this image is circulated (e.g., within a police community or via the media) with the aim of prompting identification by someone who is *familiar* with the offender. In the study, participants each viewed a photograph of a target face for a short period of time (10 seconds) with the intention of later constructing a composite. Targets were premiership-level footballers and participants were recruited as face constructors based on being *unfamiliar* with them (i.e., non-football fans). Two days later, participants described or did not describe the face using standard police interviewing procedures and constructed a feature-based composite using the PRO-fit system. Subsequently, a group of football fans (i.e., participants *familiar* with the target pool) attempted to name the composites. In this type of design, participants will more often correctly name effective composites. Frowd and Fields reported that composite likenesses were more effective in the no-description compared to the description condition, indicating the presence of a verbal overshadowing effect.

However, the no-description participants in Frowd and Field's study did engage in some form of face description. Feature-based composite systems involve large databases of individual features, and require selection of feature labels from menus to produce a manageable pool of exemplars for consideration (e.g. broad noses, large eyes). Therefore, to verify effects of verbal overshadowing, they included an additional condition, which inserted a delay (> 10 minutes) between describing the face and composite construction. The assumption was that the effect of the description would subside over time, to allow a 'release' to occur from verbal overshadowing (c.f., Meissner & Brigham, 2001). Consistent with this idea, Frowd and Fields (2011) found better performance (i.e., more effective composites) when there was a 30-minute delay compared to no delay between describing the face and constructing a composite, but that these composites did not differ in effectiveness from those produced in the no-description condition. These findings suggested that a negative effect of description (vs. no-description) had occurred in the no delay condition; also, that improved performance in the post-description delay condition was evidence of a release from verbal overshadowing.

The effects of verbal overshadowing, however, were only detectable in this study when using particularly sensitive measures of naming. It turned out that participants constructed composites that were correctly named infrequently (i.e., 1.4% mean correct by spontaneous naming), presumably due to the very short (10 second) encoding duration used. To increase naming rates and hence sensitivity in detecting changes in their interview conditions (description no-delay vs. description delay), a forced-choice naming task was administered. Here, participants knew the specific pool of targets (10 in total) to which the composites corresponded. When using this relatively more sensitive task, the verbal overshadowing effect detected was small-to-medium in size.

Weak experimental effects translate into small percentage differences in naming rates that are unlikely to impact noticeably upon detecting a perpetrator within a forensic situation. In this context, it may be that verbal overshadowing is of limited concern to policing. However, Frowd and Field's (2011) study was hampered by very low spontaneous correct naming rates for the composites that were constructed, and so proxy measures to naming were relied on to detect verbal overshadowing effects. These proxy measures are less relevant to forensic settings. An objective for the current project then was a design that would assess whether verbal overshadowing has any real-world impact. For this purpose, we utilise an experimental design that should have sufficient power to be able to detect a medium-sized effect of verbal overshadowing using a spontaneous naming task, an appropriate proxy to how composites are presented and named in the real world. Detecting a medium-sized effect [where the Odds Ratio, *OR*, or *Exp (B)* = 2.5] (Sporer & Martschuk, 2014) would indicate that the task of describing the target face immediately prior to composite construction more than halves the likelihood that the constructed face would be successfully named. This would be indicative of verbal overshadowing causing a worthwhile detriment to composite naming, substantially reducing the usefulness of composites as an investigative tool within a forensic setting.

In Experiment 1, we set out to replicate and extend Frowd and Field's (2011) findings concerning a release from verbal overshadowing: that is, we expected correct naming of composites to increase significantly following a delay between face description and composite construction. In a multi-stage study, participants described a target face and then constructed a composite using a modern 'feature' system immediately, or after 30-minutes. Given that the presence and size of verbal overshadowing effects, as applied to line-up tests, appears sensitive to changes in experimental parameters (cf., Alonga et al., 2014; Schooler, 2014) we further included varying conditions relevant to forensic composite production. We varied both the attention given to the face at encoding, and the delay between encoding and face description. We assessed the effectiveness of resulting composites for prompting recognition among another group of participants who were familiar with the targets. The outcome was an unexpected result: unlike Frowd and Fields (2011), correct naming of composites *reduced* significantly following a delay between description and construction for targets encoded two days previously. Experiment 2 recruited independent judges to examine the usefulness of the descriptions themselves, that had been generated by participants in Experiment 1, for discriminating between the target faces. This highlighted that the conditions under which the descriptions generated in Experiment 1 were more diagnostic of identity were also those conditions where a description delay had impaired feature-based composite production. Finally, in Experiment 3, a similar effect of description production was also evident when participants reproduced the face from memory one day after encoding the target using a holistic-based system. This was a second surprising result, as, unlike a feature-based system, this more recent method to create a face is designed to capitalise on face recognition rather than recall— and so does not depend on a witness having to describe a face.

EXPERIMENT 1

In Experiment 1, participants unfamiliar with our target faces constructed a single composite immediately following an interview to recall the face (the description no-delay condition), or after a 30-minute delay (the description delay condition). We then presented these composites to participants who were potentially familiar with these identities. We expected to replicate a release from verbal overshadowing (i.e., more effective composites constructed in the description delay vs. description no-delay condition) when composite production took place two days after intentionally encoding the target, conditions similar to those used by Frowd and Fields (2011). We also included two further manipulations relevant to forensic situations.

First, participants viewed a target face under intentional or incidental encoding conditions. A witness may at times be unaware that a crime is taking place (e.g., during a distraction burglary) and so he or she does not anticipate the need to later recall or recognise an offender (incidental encoding). In contrast, intentional encoding may result in a stronger memory trace of an offender. Shapiro and Penrod (1986) in their meta-analysis of studies of facial identification found both attention to the target face and knowledge of the ensuing memory test were associated with a greater number of correct identifications. We similarly anticipate that intentional encoding will lead to the production of more identifiable composites.

Second, participants provided a description either 3-4 hours or 2 days after viewing the target. Whilst a delay of one or two days is typical of forensic situations, it is not inconceivable that a witness could produce a composite on the same day as the crime. Feature-based composites are identified reasonably well when constructed up to a few hours, with reports of mean correct naming at around 16 to 25% (e.g., Brace, Pike, & Kemp, 2000; Bruce, Ness, Hancock, Newman, & Rarity, 2002; Davies, van der Willik, & Morrison, 2000; Frowd et al., 2005). However, after 2 days, composite naming is often low (≤ 5% correct; for further discussion, see meta-analysis by Frowd, Erickson, Lampinen, Skelton, McIntyre, & Hancock, 2015). Accordingly, we expected participants to construct more identifiable composites following a short (3-4 hour) versus long (2 day) post-encoding delay. As such, correct naming should be higher for faces constructed in the former time interval, increasing the chances of finding an effect, should one exist.

We also attempted to boost spontaneous naming rates of composites in two ways: by presenting target faces for encoding via video clips (vs. static photographs) and for a longer duration (30-60 seconds here vs. 10 seconds in Frowd & Fields, 2011). This allowed us to investigate the impact of face recall in situations where naming choices are considerably less constrained (cf. Frowd & Fields, 2011), a situation potentially more informative for forensic practice. Nevertheless, we anticipated that our ability to observe an effect of interview (description no-delay vs. description delay to construction) would depend upon both the effects of attention at encoding (incidental vs. intentional) and post-encoding delay. The encoding and retention manipulations may result in conditions where memory strength for the face is already likely poor (i.e., following incidental encoding and a 2 day post-encoding delay). For these conditions, the effectiveness of the composites produced may be so limited that any further detriments to performance due to describing the face immediately before composite construction (vs. following a description delay) may not be detected (i.e., due to floor effects). Therefore, we may anticipate a three-way interaction between interview, attention at encoding and post-encoding delay, and the regression-type analyses we adopt allows us to check for this as a possibility.

Method

Stage I: Composite Construction

Participants

Ninety-six students and staff from a UK based university were recruited on the basis of not following the BBC TV soap EastEnders (the targets in our study) to construct the composites (78 females, 18 males, M_{age} = 22.9 years, *SD* = 8.0, age range: 18 to 59 years). Participants received course credit or a small monetary reward for taking part. None of the participants had constructed a composite before. We randomly assigned 12 participants, with equal sampling, to each of eight experimental conditions. The appropriate University-based Ethics Committee approved experimental procedures.

Design

Participants were randomly assigned to a 2 (encoding: intentional vs. incidental) x 2 (post-encoding delay: 3-4 hours vs. 2 days) x 2 (interview: description no-delay vs. description delay) between-participants design. Ninety-six composites were constructed, one for each of 12 target faces in each of eight conditions of the experiment. The number of target faces (at least 10 per condition), and later (in Stage 2) the number of participants recruited to name the composites (at least eight per condition), were based on research (e.g., Frowd et al.,

2013) indicating that such a design should have sufficient power to be able to detect at least a medium-sized effect [*Exp(B)* = 2.5] with our planned regression-type analysis.

Materials

Stimuli consisted of six nonviolent video clips from the TV soap EastEnders, each portraying a social interaction between a male and female character for between 30 and 60 seconds. Each character appeared in only one clip, and so 12 unique characters were used as targets, six male and six female. Targets ranged from approximately 20 to 50 years of age. Composite construction took place on a PC using the feature-based system, PRO-fit (version 3.5).

Procedure

Participants individually attended the laboratory on two occasions: First, to view a video containing a target character, and second to construct a facial composite. Selection of the target was random, without replacement, within each of the eight experimental conditions.

A single experimenter (the second author) constructed the composites. This person was aware that the targets were characters from the TV soap EastEnders, but she was not privy to the specific identities of the targets (the first author having selected them). The experimenter assigned a video file and asked the participant to play the video when she had left the room. The experimenter was therefore not aware of the particular target the participant was to construct. Thus, the experimenter worked through the procedure of eliciting a description and constructing a composite based on the information each participant provided about the face that had been seen. The process was therefore participant-led: The experimenter wrote down what the participant recalled, entered this description in PRO-fit, showed features to match the description and selected and adjusted individual features as directed by the participant. This participant-led approach ensured (as far as possible) a consistent approach from the experimenter in constructing the composites across conditions.

Participants attended the laboratory on the first session and watched the assigned video clip and listened to the dialogue on headphones. Those in the incidental condition were directed to attend to the social interaction between the two characters, as they would later be asked to recall their impression of the interaction, including the dialogue. Thus, participants were not told to focus on a specific target nor about the impending composite task. Those participants in the intentional encoding condition were directed to attend closely to the facial appearance of either the female or the male character, again based on the participant's assignment, as they would later construct a facial composite of that target face. All participants reported being unfamiliar with the target for which they had been assigned.

Participants returned for a second session either 3-4 hours or 2 days later. The experimenter explained that they would now describe the face of either the female or the male character, as appropriate, and construct a composite. The experimenter used interviewing techniques designed to support witness retrieval (as is standard in a forensic setting) to elicit a description of the target face (e.g., Geiselman, Fisher, MacKinnon, & Holland, 1986). Participants were asked to form a visual image of the face, thinking back to the context in which they saw the face, and to recall freely the face in as much detail as possible. Subsequently, as part of a cued-recall procedure, the experimenter repeated back the participant's initial description of each facial feature (in the order: face shape, hair, eyebrows, eyes, nose, mouth and ears) and prompted participants to see if further information could be recalled about the face. According to assignment, composite construction was conducted immediately after the description (i.e., description no-delay) or after a 30-minute delay (description delay). Those assigned to the description delay condition undertook the same comprehension test used by Frowd and Fields (2011) to occupy their time.

To construct the composite, the experimenter entered the participant’s description into PRO-fit to locate approximately 20 examples per feature, and prepare an “initial” composite, a face whose appearance matched the description. The participant then guided the experimenter to exchange features with other examples, and to size, position and adjust features (e.g., changing the brightness and contrast) until the best possible likeness had been reached. Finally, an offer was made to use an artwork package to enhance the composite (e.g., by adding wrinkles or stubble). Composites took approximately fifty minutes to construct.

Stage II: Composite Evaluation

The composites constructed in Stage I were evaluated for their effectiveness using two tasks. First, a new set of participants, this time reporting to be familiar with the targets attempted to name the composites. Second, we asked a different group of participants to provide an assessment of likeness. Here, participants unfamiliar with the target characters rated each composite face based on its perceived match to a photograph of the intended target. Likeness ratings typically function as a good proxy to composite naming (Frowd, Bruce, McIntyre, & Hancock, 2007a) and we expected similar outcomes across both measures.

Composite Naming

Participants

Sixty-four volunteers from two UK universities and their local surrounding areas took part, recruited on the basis that they reported being *regular viewers* of EastEnders (49 females, 13 males, we did not record gender for two participants, M_{age} = 25.8 years, *SD* =

11.1, age range: 18 to 67 years). We randomly assigned participants with equal sampling to view one of eight sets of composites as per the experimental design.

Materials

Composites were printed in greyscale, the image modality for this composite system, one per page (10 x 15 cm). Figure 1 presents example composites. There were eight composite sets, each set including the 12 composites from a single condition. Colour photographs showing head and shoulder front-on views of the 12 targets were also printed, one per page (10 x 15cm).

Figure 1 about here

Design and procedure

Participants viewed composites created in Stage I that belonged to a single experimental condition, and thus no participant viewed more than one composite belonging to the same target identity. Thus a 2 (encoding condition) x 2 (post-encoding delay) x 2 (interview) between-participants design was used. Participants carried out this self-paced task, which took approximately 10 minutes, individually. They viewed the 12 composites from a single set one at a time (in a different random order for each participant) and attempted to provide any name that came to mind (real or stage), or gave a ‘don’t know’ response. Following this, to verify familiarity with the corresponding targets, participants were asked to name each of the 12 photographs of the targets, presented likewise.

Results

Participants correctly named between 9 and 12 target photographs (*M* > 88.5% and *SD* < 1.3% correct in each cell of the design), indicating very good familiarity with the relevant target photographs. Responses to composites were coded as correct (assigned a numeric value of 1) when participants gave an appropriate name, or incorrect (value of 0) when no name or a mistaken (wrong) name was given. The inclusion of a naming response within the following analyses was conditional upon the participant having correctly named the relevant target photograph: composites associated with photographs of targets unknown by participants were unlikely to attract a correct name and so such responses were treated as missing data. Mean correct naming across the 96 composites (*M* = 11.9% overall) was worse than that obtained for the target photographs. This is a typical outcome given that, unlike photographs, composites do not represent a veridical image of the person, thus making the face more difficult to recognise. To supplement these data, in a separate analysis presented later in this section, participant responses were re-scored in terms of mistaken names given to provide a further measure of composite effectiveness.

Within our experimental design, although each participant attempted to name each of 12 composites, no participant named more than one composite belonging to the same target. Thus by design, observations by-item were independent. However, whilst naming scores contributed from the same participant do not always correlate, in some instances, there may be some degree of dependency. Generalized Estimating Equations (GEE) can be used to fit a binary logistic regression to model naming scores (as a dichotomous dependent variable, DV; correct vs. incorrect) and their relationship to manipulated variables (predictors) while accounting for the possibility of dependency within each participant's set of naming responses. GEE provides a combined by-participants and by-items model, and as such is more statistically powerful than ANOVA (Ballinger, 2004).

Predictors were encoding (coded as 1 = incidental, 2 = intentional), post-encoding delay (1 = 3-4 hours, 2 = 2 days) and interview (1 = description no-delay, 2 = description delay); throughout the analyses, the lowest-coded category was selected as the reference, and reported Beta (*B*) coefficients reflect this scheme. To model within-participant correlations in responses we applied two working correlation structures (independent vs. exchangeable) to a saturated regression model (i.e., including all predictors). We took as optimal the correlation structure that gave rise to the smallest QIC value (Quasi likelihood under Independence Models Criterion; Pan, 2001; see Cui & Qian, 2007). This led to the application of an exchangeable correlation matrix. This seems appropriate as within-participant naming responses may be expected to correlate equally over time, with a participant's earlier naming responses not expected to influence their later responses[1].

First, we built a saturated model that included all main variables and interactions. This allowed us to test our predictions related to encoding, post-encoding delay and post-interview delay and their three-way interaction (as outlined in the Introduction). We retained the main variables relating to our key manipulations within the final model to estimate the unique contribution of each to the prediction of naming rates. Interactions between variables were tested for inclusion using the backward elimination method beginning with sequential removal of interactions contributing least to the variance within naming (based on the established standard criteria of $p > .1$ and lowest Wald X^2). This method is sensitive to detecting variables whose influence is determined by the presence of other variables (suppressor variables), and therefore seems appropriate for testing for an effect of verbal overshadowing, given that this phenomenon has been found to be sensitive to changes in experimental parameters (cf., Alonga et al., 2014; Schooler, 2014).

When built, standard errors (SE) of Beta (*B*) coefficients were inspected for markers of model instability (of which none were observed). Analyses were carried out with SPSS version 22 (IBM SPSS Statistics 22, Chicago, IL). For clarity, throughout, we report only main effects and interactions included within the final regression model. Terms used in the presented GEE equations include: Wald X^2 and associated degrees of freedom *(df)* and *p*-value; *SE(B)*, the standard error (variability) of the predictor's Beta *(B)* coefficient; *Exp(B),* the Odds Ratio[2] (a measure of effect size); and (in brackets) 95 percent CIs for *Exp(B).*

Correct naming: Correct naming of composites was our primary DV. Correct naming responses are summarized in Table 1 - see *Note* for how these values were calculated.

Table 1 about here.

Correct naming rates for the encoding variable were in the predicted direction, with more identifiable composites produced following intentional (13.3%) compared to incidental (10.45%) encoding conditions. However, although retained within the final model, encoding was not found to be a significant predictor [$X^2(1) = 1.50$, $p = .22$]. Post-encoding delay was a significant predictor [$X^2(1) = 7.27$, $p = .007$], since, as predicted, correct naming scores were lower following a post-encoding delay of 2 days (9.33%) compared to 3-4 hours (14.4%). Post-interview delay was also a significant predictor [$X^2(1) = 10.82$, $p = .001$]. However, contrary to our prediction based on Frowd and Fields's data (2011), correct naming was *lower* when a 30-minute delay (8.3%) compared to no-delay (15.2%) occurred between face description and construction. Finally, the above main effects were qualified by a significant interaction between post-encoding x post-interview delay [$X^2(1) = 4.41$, $p = .036$].

As can be seen from Table 2 the interaction appears to arise as, compared to all three other conditions, the rate of correct naming was significantly *lower* for the 2 days, description delay condition, a deficit which in all cases was accompanied by a medium effect size [*Exp (B)* > 2.5]. Compared to this condition, correct naming rates were more than doubled in the 2 days, description no-delay condition [*B* = 1.31, *SE(B)* = 0.38, *df* = 1, *p* = .001, *Exp(B)* = 3.71 (1.75, 7.88)], 3-4 hours, description delay condition [*B* = 1.17, *SE(B)* = 0.40, *df* = 1, *p* = .003, *Exp(B)* = 3.21 (1.48, 6.98)], and 3-4 hours, description no-delay condition [*B* = 1.46, *SE(B)* = 0.35, *df* = 1, *p* < .001, *Exp(B)* = 4.29 (2.15, 8.58)]. Note, at the shorter (3-4 hour) post-encoding delay, the description no-delay and description delay conditions did not significantly differ [*B* = 0.29, *SE(B)* = 0.30, *df* = 1, *p* = *.33*, *Exp(B)* = 1.34 (0.74, 2.40)]. These findings imply that following a longer post-encoding delay (2 days), inserting a 30-minute delay after providing a description, is detrimental to composite construction.

Table 2 about here.

Mistaken naming: We now consider an analysis of mistaken naming: when a composite gives rise to a wrong name, in comparison to a 'don't know' response. This measure provides further information on composite quality, as an increase in mistaken (i.e., wrong) names indicate a less accurate composite. From our overall dataset, we removed correct responses and (as above) any composite data points where the corresponding target photograph had not been correctly named. We then calculated the number of mistaken names as a proportion of the total number of remaining responses (out of 654 responses in total for mistaken, coded as 1, and 'don't know', coded as 0); as such, mistaken naming is independent of correct naming. Although, this measure may be less precise when the number of correct names is high (as there would not be many instances left with which to establish incorrect naming), this issue should not be relevant overall, as typically correct naming of composites does not reach ceiling levels (cf. Frowd et al., 2015 meta-analysis). Here, mistaken naming made up 40.5% of all incorrect responses.

Mistaken naming responses are summarized in Table 3. A GEE procedure was applied as outlined above. The final model retained three main effects, all non-significant (see *Note* for Table 3) indicating that mistaken naming rates did not differ as a function of the experimental manipulations.

Table 3 about here.

Composite Likeness Ratings

Participants

Fifty-two participants from a UK university and its local surrounding area volunteered on the basis of being *unfamiliar* with the TV Soap EastEnders (20 males, 32 females, M_{age} = 21.59 years, *SD* = 7.15, age range 18 to 52 years). We randomly assigned participants with equal sampling to view one of two sets of target photograph-composite pairs (26 participants per set).

Materials

We printed each composite and its corresponding colour target photograph on a single A4 page (one pair per page). Photographs appeared on the left and composites on the right, each sized to approximately 10cm wide x 12cm high. Two sets of target photograph-composite pairs were prepared, each including the 48 composites constructed under either incidental or intentional encoding conditions.

Design and procedure

To avoid lengthy testing sessions, participants were randomly assigned to rate target photograph-composite pairs from one encoding condition (incidental or intentional; between-participants), but from both post-encoding delay conditions (3-4 hours and 2 days; within-participants) and interview conditions (description no-delay and description delay; within-participants). Participants carried out this self-paced task, which took approximately 15 minutes, individually. We randomised the order of presentation of target photograph-composite pairs for each participant. Participants considered each pair in turn and rated the likeness of the composite to the face in the photograph on a scale of 1 (poor likeness) to 10 (good likeness).

Results

The rating data (Table 4) were analysed using a 2 (encoding) x 2 (post-encoding delay) x 2 (interview) Mixed-Factorial ANOVA, with encoding condition as a between-participants factor. There was a main effect of interview, $F(1,50) = 17.27$, $p < .001$, $MSe = 0.24$, $\eta_p^2 = .26$, qualified by a significant interaction with encoding, $F(1,50) = 12.07$, $p = .001$, $MSe = 0.24$, $\eta_p^2 = .19$. The main effect of interview was similarly significant in the analysis by-items, $F(1,11) = 5.08$, $p =.046$, $MSe = 0.38$, $\eta_p^2 = 0.32$, and the encoding condition x interview interaction marginally significant, $F(1,11) = 3.74$, $p = .079$, $MSe = 0.36$, $\eta_p^2 = .25$. All other main effects and interactions were non-significant. Follow-up pairwise comparisons by-participants revealed that for targets viewed under intentional encoding conditions, participants rated composites as worse likenesses when they had been constructed following a description delay ($M = 3.02$, $SD = 1.20$) compared to no-delay ($M = 3.53$, $SD = 1.18$), $F(1,25) = 26.71$, $p < .001$, $MSe = 0.13$, $\eta_p^2 = .52$. For targets viewed under incidental encoding conditions, no differences arose in ratings between the description delay ($M = 3.29$, $SD = 1.49$) and description no-delay conditions ($M = 3.34$, $SD = 1.60$, $p = .62$). No differences in ratings arose between composites constructed under incidental versus intentional encoding conditions within either the description no-delay (*p* = .62) or description delay (*p* = .47) conditions.

Table 4 about here.

Discussion

Our key finding was that inserting a 30-minute delay between face description and composite construction impaired composite effectiveness for some experimental parameters. Specifically, there was a reduction in correct naming (mistaken naming rates did not differ). This outcome was evident when construction followed a delay of 2 days, a usual timeframe experienced by witnesses in real-world situations. Likeness ratings confirmed the detrimental effect of a post-description delay following intentional encoding of the target face.

We note that findings related to post-encoding delay and encoding conditions did not consistently arise across both naming and likeness measures. Likeness ratings are usually a good proxy to naming, and scores here from the two measures were positively correlated [*r(94)* = .22, *p* = .032]. Nevertheless, there may have been differences in the sensitivity of both tasks to detecting effects of our manipulations. For example, unlike naming participants, those undertaking the likeness-rating task were unfamiliar with the relevant target faces. External features (i.e., face shape, ears and hair) are of greater importance to unfamiliar relative to familiar face recognition (e.g., Ellis, Shepherd, & Davies, 1979). Thus, external features may have attracted more attention when composites were rated for likeness to a corresponding target compared to when participants attempted to give a name. We discuss findings in more detail arising from the naming and rating tasks, and their implications, in the General Discussion.

Our findings regarding the effects of description delay contrast with those of Frowd and Fields (2011). This was despite both experiments using the same feature-based composite system (PRO-fit), intentional encoding and a 2-day post-encoding delay. Under these conditions, they instead found evidence for a beneficial effect of inserting a 30-minute delay between description and composite construction (vs. no-delay), although spontaneous naming was poor, with only 1.4% correct names generated. Our naming rates were substantially higher than those obtained by Frowd and Fields (2011). Indeed, when composites were constructed immediately after a description (i.e., the description no-delay condition), we found correct naming rates at 3-4 hour (16.1%) and 2 day post-encoding delays (14.3%) to be equivalent (although the means trended in the predicted direction). Unexpectedly this indicates little detriment of longer retention intervals upon composite effectiveness.

Further, contrary to predictions, encoding the target under incidental compared to intentional encoding conditions did not reduce correct naming. Previous research has shown that participants viewing videos of unfamiliar target faces that display both full-face and three-quarter-views (vs. full-face views only) produce more effective PRO-fit composite likenesses (Ness, Hancock, Bowie, Bruce, & Pike, 2015). By making use of videos depicting multiple viewpoints (compared to static photographs, as used by Frowd & Fields, 2011), we may have provided participants with more information to draw upon when later constructing their face composites. Additionally, we may expect the longer encoding duration used (30-60 seconds vs. 10 seconds) to be associated with an increase in correct naming of composites, similar to identification of real (non-composite) faces (e.g., Shapiro & Penrod, 1986).

The key result arising from Experiment 1 implies that the process of describing a face can be useful to producing a feature-based composite; under some circumstances, reducing access to this description in memory by inserting a delay post-description (here, a delay of 30 minutes) hinders effective composite construction. One possibility is that under these particular circumstances, the generated description contains useful details about the face, which a witness subsequently relies upon for composite construction. Inserting a delay between describing the face and composite construction may make it more difficult for a witness to utilise information from their description effectively—for example, witnesses may simply forget facial information during this time. On this account, we may expect the contents of the descriptions generated in Experiment 1 to vary according to the conditions under which participants encoded the target (intentional or incidental) and subsequently provided their description (whether 3-4 hours or 2 days post-encoding). More specifically, we may expect that those conditions where participants demonstrated a detrimental effect of a (30-minute) description delay on composite construction (i.e., 2 days post-encoding and/or intentional encoding) to produce descriptions that were in some way *more* useful for discriminating the target from other faces. While intentional encoding would be expected to promote a more effective description, this hypothesis also implies that a description is more effective after a 2-day (cf. 3-4 hr) delay, perhaps as confidence reduces for less important information at longer intervals of time (see General Discussion). The assumption is also that, under these (intentional encoding / 2-day delay) conditions, inserting a post-description delay may have interfered with the ability of participants to utilise informative aspects of their descriptions. We explore this possibility in Experiment 2.

EXPERIMENT 2

We aimed to evaluate the content of the participants' facial descriptions produced in Experiment 1. Here, we asked 'independent judges' (i.e., participants who had not been exposed to the target videos used in Experiment 1) to attempt to match the descriptions to photographs of their corresponding targets (a communication-accuracy procedure; see Brown & Lloyd-Jones, 2003; Fallshore & Schooler, 1995; Malpass, Lavigueur, & Weldon, 1973). This procedure allowed us to assess conditions (i.e., incidental vs. intentional encoding, and 3-4 hours vs. 2 days post-encoding delay) under which descriptions were more discriminating of their targets. Descriptions contain information from both free recall and cued recall, and so we examined whether one or both of these components of the description would be more or less useful. Prompts for more information (i.e., via cued recall) can elicit details over and above free recall, but may also lower participants' criterion for reporting information, and this is turn could lead to an increase in incorrect descriptors (cf. Koriat & Goldsmith, 1996; Meissner et al., 2001).

Additionally, in a separate analysis, two independent coders coded the quality of the descriptions participants generated in Experiment 1 in terms of the number and accuracy of the descriptors provided.

Matching descriptions to photographs of target faces

Participants

Ninety-six students and staff from a UK University were recruited as independent judges on the basis that they reported being *unfamiliar* with the TV soap EastEnders (34 males, 62 females, M_{age} = 30.33 years, *SD* = 12.18, age range 18 to 68 years). None had previously taken part in Experiment 1. We randomly assigned independent judges, with equal sampling, to one of 16 conditions. All received a small monetary reward. The appropriate University-based Ethics Committee approved the experimental procedures.

Materials

We took the 96 descriptions of target faces generated by Stage I participants in Experiment 1: One description for each of the 96 participants that constructed a composite. For each description, we derived two versions, one including free recall only and one including free and cued recall. This gave 192 descriptions in total (two versions of each of the descriptions generated by the 96 Stage I participants in Experiment 1). Each description was typed and presented on an individual card in a standardised format. Information appeared under the following headings: gender, overall appearance, face shape, hair, eyebrows, eyes, nose, mouth, ears and other information. The 12 colour target photographs used in Experiment 1 were also required (each photograph depicted a unique target person).

Design

We sub-divided each of the two sets of 96 descriptions (set 1, free recall; and set 2, free and cued recall) into eight sub-sets relating to the eight experimental conditions under which Stage I participants in Experiment 1 originally generated the descriptions. In total, this gave 16 separate sub-sets of descriptions forming a 2 (encoding: incidental vs. intentional) x 2 (post-encoding delay: 3-4 hours vs. 2 days) x 2 (interview: description no-delay vs. description delay) x 2 (recall content: free vs. free and cued) between-participants design. The dependent variable was accuracy in matching a description to its corresponding target photograph (correct or incorrect). This meant that each sub-set included 12 descriptions, consisting of one description corresponding to each of the 12 unique target photographs. In this way, the matching task was designed so that no independent judge matched more than one description belonging to the same target photograph. Six judges viewed each of the 16 description sub-sets. For each sub-set, we generated three different random orders of presentation, with each order shown to two independent judges.

Procedure

Independent judges carried out this self-paced matching task individually, taking approximately 25 minutes to complete. We presented the 12 target photographs on a table, and asked each independent judge to read a description twice before laying it face down in front of the target photograph they believed matched the description. Judges were asked to make each matching decision independently of other decisions and therefore they could match more than one description to a single target photograph. Each independent judge repeated this procedure for all 12 descriptions.

Results

No independent judge matched more than one description belonging to the same target, meaning observations by-items were independent. However, as for the naming analysis within Experiment 1, judges each contributed multiple responses and so responses from the same judge may not necessarily be uncorrelated. To model the potential for non-independence within the data we again used GEE to fit a binary logistic regression model to our dichotomous DV (1, correct match; 0, incorrect match).

We did not expect matching decisions made by the independent judges to differ systematically for sets of descriptions belonging to the two different interview conditions (description no-delay vs. description delay). This is because in Experiment 1, all descriptions were collected from Stage I participants before they undertook the interview condition manipulation. A GEE regression model including interview as a single categorical predictor (coded as 1 = description no-delay, 2 = description delay) confirmed that there was no effect of this predictor on the accuracy of matching decisions [$X^2(1) = 0.81$, $p = .37$], and so, the manipulation, interview, was not included as a predictor in the analysis that follows.

The percentage of correct naming responses are summarized in Table 5. We proceeded with the GEE analysis using the method described for Experiment 1. As before, we retained the main variables relating to our important manipulations for estimating their unique contributions to the prediction of naming rates, and interactions between these variables were tested for inclusion within the final model using the backward elimination method (with sequential removal when $p > .1$ and lowest Wald X^2). The predictors were encoding (1 = incidental, 2 = intentional), post-encoding delay (1 = 3-4 hours, 2 = 2 days) and recall type (1 = free recall only, 2 = free recall and cued recall); as before, the lowest coded category was selected as reference for comparison.

The final regression model revealed that recall type was not a significant predictor [$X^2(1) < .001$, $p = 1.00$], thus cued recall did not add further distinguishing information over and above free recall. Encoding was a significant predictor [$X^2(1) = 16.48$, $p < .001$], as expected, such that descriptions were better matched with photographs viewed under intentional (56.94% correct matches) compared to incidental (44.10% correct matches) encoding conditions [*B* = 0.52, *SE(B)* = 0.13, *Exp(B)* = 1.68 (1.31, 2.16)]. Post-encoding delay was also a significant predictor [$X^2(1) = 4.89$, $p = .027$]. Here, descriptions emerged better matched with photographs when elicited under the longer (2 days, 54% correct matches) compared to the shorter (3-4 hours, 47.05% correct matches) post-encoding delay [*B* = 0.28, *SE(B)* = 0.13, *Exp(B)* = 1.33 (1.03, 1.71)].

Table 5 about here

Description Quality

We next assessed the number of descriptors generated by each participant in Experiment 1 (Stage I: construction) and their accuracy. An independent coder checked the 92 descriptions against a corresponding coding protocol generated for each face (based on modal facial descriptors elicited from eight independent participants[3]). Correct descriptors matched the protocol and incorrect descriptors did not. Subjective details referred to nonspecific facial features, such as personality impressions (e.g., kind, mean). For the sake of brevity, we report significant outcomes ($p < .05$) only.

A 2 (encoding) x 2 (post-encoding delay) between-participants ANOVA was carried out on the total number of details recalled about the face (i.e., the number of correct + incorrect + subjective details). This revealed more details were recalled under intentional (*M* = 33.92, *SD* = 9.02) compared to incidental (*M* = 27.48, *SD* = 7.33) encoding, *F*(1,92) = 15.67, $p < .001$, *MSe* = 63.47, $\eta_p^2 = .15$, and following shorter (3-4 hours, *M* = 32.98, *SD* = 8.11) compared to longer (2 days, *M* = 28.42, *SD* = 8.94) post-encoding delay, *F*(1,92) = 7.87, $p = .006$, *MSe* = 63.47, $\eta_p^2 = .079$.

A further 2 (encoding type) x 2 (post-encoding delay) between-participants ANOVA was carried out first on the number of correct and then incorrect details recalled about the face. Significantly more correct details were recalled under intentional (*M* = 12.40, *SD* = 3.93) compared to incidental (*M* = 10.00, *SD* = 3.61) encoding, *F*(1,92) = 9.59, $p = .003$, *MSe* = 14.37, $\eta_p^2 = .09$. Significantly fewer incorrect details were recalled in the longer, 2 day (*M* = 2.85, *SD* = 1.58) compared to the shorter, 3-4 hours (*M* = 3.77, *SD* = 2.13) post-encoding delay, *F*(1,92) = 5.73, $p = .019$, *MSe* = 3.52, $\eta_p^2 = .06$.

Discussion

Experiment 2 evaluated the content of the descriptions Stage I participants in Experiment 1 generated prior to constructing their composites. We found that descriptions were more effective in distinguishing among targets when generated following intentional (cf. incidental) encoding of the target face. However, we also found descriptions emerged less effective when recalled after 3-4 hours (cf. 2 days). When taking these results together with our analysis of composite naming in Experiment 1, it seems that descriptions were most useful (i.e. better matched to their corresponding target by independent judges) when they had been generated under those same conditions that showed a detrimental influence of a (30-minute) description delay on composite effectiveness (i.e., the intentional encoding conditions and 2 day post-encoding delay conditions). This implies that in these conditions, Stage I participants in Experiment 1 were relying on their descriptions in memory when constructing their composites, and that reducing access to these descriptions (via a 30-minutes delay) interfered with composite effectiveness.

Descriptions also contained a greater amount of information when generated under conditions where a stronger memory trace for a target face was expected. Previous research in forensic settings has also found eyewitnesses experiencing shorter delays between viewing the event and attempting recall give more information (Ellis, Shepherd, & Davies, 1980; Penrod, Loftus, & Winkler, 1982; Turtle & Yuille, 1994; van Koppen & Lochun, 1997). Further, intentional compared to incidental encoding has been found to lead to an increase in participants' recall of information about an event (Davies & Hine, 2007; Yarmey, 2004). Nevertheless, increased description quantity was not consistently associated with more discriminating descriptions. Whilst descriptions generated in the shorter (3-4 hour) post-encoding delay condition contained more descriptors, this condition elicited poorer description-to-target matching (vs. 2 day post-encoding delay). More generally, we found a weak (and non-significant) trend indicating that the greater the number of face descriptors (incorrect + correct + subjective descriptors) the less useful the descriptions were for distinguishing the target ($r(94) = -.18$, $p = .082$). These findings imply that the usefulness of a description is not necessarily a function of the quantity of information it contains. In keeping with this, Ellis et al. (1980) found that although recall at one day versus one hour following the encoding of a target face led to fewer descriptors about that face, independent judges (like those used here) matched both descriptions to their target face at a similar rate.

Instead, the accuracy of descriptions appears to be important. Those conditions eliciting more useful (i.e., more discriminating) descriptions also generated descriptions containing a greater amount of accurate information, either in terms of more correct details (under intentional vs. incidental encoding conditions) or fewer erroneous details (following the 2 day vs. 3-4 hr post-encoding delay). Further, overall, the percentage of descriptions correctly matched to their corresponding target was positively correlated with description accuracy (incorrect / (incorrect + correct descriptors), $r(94) = .21$, $p < .05$).

The findings from Experiment 2 imply a role for the contents of the description, particularly its accuracy, in contributing to the effectiveness of the composite produced. The description delay conditions that showed a reduction in composite effectiveness in Experiment 1 were also those conditions that elicited more useful (i.e., discriminating) descriptions as demonstrated in Experiment 2. Thus our findings so far indicate that, at least for modern feature-based systems, under some conditions, a description proves useful for face construction.

Recent developments in composite construction, however, have seen the emergence of alternatives to modern feature-based systems. Unlike a feature-based system, these more recent methods have been designed to capitalise on face recognition rather than recall ability - and therefore do not depend on a witness having to describe the face. For this reason, the effect of producing a description of the target face on subsequent composite construction may lead to different outcomes using these newer ‘holistic’ systems—in particular, by potentially not impacting on composite accuracy following inclusion of a 30-minute delay after face recall. Experiment 3 addresses this possibility.

EXPERIMENT 3

‘Holistic’ systems involve selection and ‘breeding’ of whole faces (e.g., E-FIT-V or EFIT-6, EvoFIT, ID; Frowd et al., 2010; Tredoux, Nunez, Oxtoby, & Prag, 2006; Valentine, Davis, Thorner, Solomon, & Gibson, 2010; see Frowd, 2017 for a detailed review). It is assumed that this process is more closely aligned to the way in which we naturally perceive and recognise faces: as whole entities, with individual features perceived in the context of other features and their overall spatial configuration (Frowd, Hancock, & Carson, 2004; Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Notably, selection of whole faces within holistic-based systems can proceed ergonomically in the absence of a description. We took advantage of this situation to conduct a formal test of the effect of verbal overshadowing on the effectiveness of face construction. Experiment 1 did not include a condition whereby face construction took place in the absence of a description, as omitting face recall would have made face construction difficult for a feature-based system (Frowd et al., 2005). Given that we are using a holistic-based system, we anticipated verbal overshadowing would now be evident. This is because, holistic systems place potentially greater emphasis upon recognising faces as whole entities, a notion not fully capitalised in modern feature-based composite systems.

In Experiment 3, Stage I participants viewed a video clip of an unfamiliar target face and 22-26 hours later described the target or did not (in a no-description condition) and then constructed a composite using one of the holistic-based systems, EvoFIT. We now expected to observe an effect of verbal overshadowing, with less effective composites occurring in the description than no-description condition (i.e., shown by a reduction in correct naming by Stage II participants). The same as for Experiment 1, we included an additional condition where participants described the target and then constructed a composite after a 30-minute delay. We anticipated that inserting a delay between face recall and composite construction would allow for a “release” from verbal overshadowing, as the effect of description would subside over time. We therefore anticipated improved naming rates for the description delay compared to description no-delay condition.

Method

Stage I: Composite Construction

Participants

An opportunity sample was recruited consisting of 40 staff and student volunteers from a UK university (20 male, M_{age} = 20.60 years, *SD* = 1.80, age range 18 to 24 years). None of the participants had constructed a composite before. The appropriate University-based Ethics Committee approved experimental procedures.

Design

Participants were randomly assigned, with equal sampling, to one of four interview conditions, although we only report the outcome for three conditions here, description no-delay, description delay, and no-description[4]. Thirty composites were constructed, one for each of 10 targets in each of these three conditions in a between-participants design.

Materials

Target stimuli were 10 non-violent video clips, each portraying a different member of staff (5 male, 5 female) from a retail outlet in NW England giving directions to a local town centre. Age of the staff ranged from approximately 20 to 50 years, and video clips were about 30 seconds in length. A PC was used with EvoFIT software version 1.3.

Procedure

As for Experiment 1, participants attended the laboratory individually on two separate occasions, this time 22-26 hours apart. As before, a participant-led process was undertaken. The experimenter (different to Experiment 1) was not privy to the targets and controlled the software to construct the face utilising information each participant provided. On attending the first session, participants watched a video clip, with explicit instructions to pay attention to the person therein and what he or she said. Participants viewed one of 10 video clips, randomly selected without replacement, within each of the three experimental conditions. Participants listened to the dialogue on headphones. All participants reported that the identity of the person in the video was unfamiliar to them. In the second session, two groups of participants were asked to describe the target face in an interview, which made use of techniques designed to support witness retrieval (as previously described for Experiment 1). The composite was constructed using EvoFIT after providing this description (description no-delay), or after a 30-minute delay (description delay). A third group did not provide a description (no-description) prior to composite construction.

There have been various iterations of EvoFIT (see Frowd, 2017), but the version used here would seem to be fairly representative of a holistic system. We followed the procedure as described in detail in Frowd et al. (2010). In brief, external features (hair, forehead, ears and neck), when selected by a participant at the start, were blurred (via applying a Gaussian filter) and shown in subsequent face arrays along with internal features (the region including eyes, brows, nose and mouth). Participants were asked to select best overall matches to their given target from arrays which presented faces that changed first by facial shape (shape and position of features on the face), then facial texture (greyscale colouring) and finally by both shape and texture. Characteristics of selected faces were combined using an “evolutionary” algorithm, and the process was repeated until the participant believed that the best likeness had been achieved. For the final array of faces, participants were asked to choose a single item that most accurately represented the target face. For this face, external features were restored (i.e., made fully visible) and software tools were introduced to allow the person to improve the likeness, first by enhancing holistic properties of the face, including its perceived extroversion, attractiveness, masculinity and health; and then by adjusting the size and placement of individual features. Composites took approximately an hour to construct.

Stage II: Composite Evaluation

Naming

Participants

An opportunity sample of 48 members of staff volunteered, participants who worked in the retail outlet at which the targets had been filmed (18 males and 30 females, M_{age} = 27.25 years, *SD* = 8.12, age range: 17 to 49 years).

Materials

The composites were printed in greyscale, the image modality of this production system, one per page (8.5 cm wide x 11.0 cm high; see examples presented in Figure 2). To make the task more realistic, and to limit guessing, the experimenter constructed two male and two female ‘foil’ composites of adult faces with EvoFIT using the construction procedure described above. The foils were of faces of similar age and general appearance to the targets, but these faces were unknown to those participants taking part in the naming study (and they were not from the target set). A colour photograph showing a head and shoulder front-on view of each of the 10 targets was also printed, one per page (12.0 x 16.0 cm).

Figure 2 about here

Design and procedure

Participants were randomly assigned, with equal sampling, to one of four experimental conditions. However, as described previously, only the three relevant conditions are reported here. Participants were invited to name 10 composites produced in Stage I that belonged to one of the three interview conditions; a between-participants design. Included in each condition were two male and two female ‘foil’ composites.

Participants completed the self-paced task individually, which took approximately 15 minutes. They were instructed that a set of composites would be seen, some of whom were of colleagues of theirs who worked in same the retail outlet. Participants thus provided a name for each composite or responded with “don’t know”. Composites (of targets and foils) were presented sequentially in a different random order for each person. Afterwards, participants were likewise presented with each of the 10 target photographs to name.

Results

All participants correctly named all 10 of the target photographs, indicating all of the targets were familiar. Therefore, all naming responses were included in the analysis. As for Experiment 1, responses to composites were coded as correct (a numeric value of 1) when participants gave an appropriate name, or incorrect (value of 0) when no name or a mistaken (wrong) name was given. Mean correct naming for the 30 composites was fairly good (*M* = 18.6% overall), appropriate for this version of EvoFIT (Frowd et al., 2015). Applying the same procedure as for Experiment 1, both correct and mistaken naming responses were analysed using Generalized Estimating Equations (GEE) to fit a binary logistic regression model. For this analysis, the regression model included the single categorical variable: interview (coded as 0 = description no-delay, 1 = no-description, 2 = description delay).

Correct Naming: The resulting model was significant for interview [$X^2(2) = 12.22$, p = .002, Table 6]. First, the effects of the presence (description conditions) versus absence (no-description condition) were tested on rates of correct naming. Correct naming scores of composites produced in the description no-delay condition did not differ from those obtained by composites made in the no-description condition [*B* = 0.05, *SE(B)* = 0.22, *df* = 1, *p* = .82, *Exp(B)* = 1.05 (0.69, 1.6)]; thus, there was no evidence of verbal overshadowing. Moreover, a release from verbal overshadowing was not observed: instead, the same as in Experiment 1, with a feature-based composite system, correct naming was significantly *lower* when a 30-minute delay (compared to no-delay) was inserted between giving a description and constructing a composite [*B* = -0.79, *SE(B)* = 0.26, *df* = 1, *p* = .003, 1/*Exp(B)* = 2.20 (1.31, 3.69)]. Correct naming scores in the description delay condition were also significantly lower than those obtained for composites produced in the no-description condition [*B* = -0.74, *SE(B)* = 0.23, *df* = 1, *p* = .001, 1/*Exp(B)* = 2.09 (1.35, 3.26)].

Table 6 about here

Mistaken Naming: As in Experiment 1, the number of mistaken names was considered relative to the total number of incorrect names (1 = mistaken, 0 = ‘don’t know’ responses; again, correct names were treated as missing data). Mistaken naming comprised 68.9% of incorrect responses. We used GEE to fit a binary logistic regression model that included the categorical variable, interview. Interview was a significant predictor [$X^2(2)$ = 12.98, *p* = .002, Table 7].

We also tested the effects of the presence (description conditions) versus absence (no-description condition) upon the likelihood mistaken names were generated. Significantly, more mistaken naming scores occurred in the no-description condition than in the description delay condition [*B* = 1.10, *SE(B)* = 0.32, *df* = 1, *p* = .001, *Exp(B)* = 3.00 (1.61, 5.58)]. Table 7 shows that the no-description condition also generated more mistaken names than the description no-delay condition, but this difference was not statistically reliable [*B* = 0.65, *SE(B)* = 0.41, *df* = 1, *p* = .11, *Exp(B)* = 1.92 (0.86, 4.27)]. As in Experiment 1, mistaken naming did not differ significantly between the description no-delay and description delay condition [(*B* = -0.45, *SE(B)* = 0.31, *df* = 1, *p* = .14, *Exp(B)* = 1.56 (0.86, 2.85)].

However, the percentage of mistaken naming for the description no-delay condition clearly falls midway between the other two interview categories, and this seems to indicate that the presence of a description seems to be having some effect. To test this assertion, we applied polynomial contrasts to test the magnitude of differences in mistaken naming rates between the no-description, description no-delay and description delay interviews (conditions entered in this order). The analysis emerged with polynomial contrasts reliable as a linear [$X^2(1) = 17.39$, $p < .001$], but not as a quadratic trend [$X^2(1) = 0.02$, $p = .89$]. In sum, naming rates for composites constructed in the description no-delay condition fell midway between the other two interview conditions, indicating that mistaken naming was highest for the no-description condition, lower for the description no-delay condition, and lower again (by the same amount) for the description delay condition.

Table 7 about here

Discussion

Using a holistic system enabled a good test of whether presence of a description negatively affects composite construction when compared to when participants engaged in no-description. There was no reliable difference in correct naming for the description no-delay and no-description conditions, with mean correct naming being virtually identical.

Thus, contrary to expectation, when using a holistic system, providing a description immediately prior to composite construction did not lead to a detriment in composite effectiveness. In fact, the analysis of mistaken names revealed a slight superiority for the description no-delay than no-description condition, as mistaken names were less frequent. Thus composites produced in the presence (compared to absence) of a description, although not correctly named any more often, were actually more accurate, as they were less easily mistakenly confused with other identities. This finding is consistent with the notion that eliciting a description allows a person to construct a more effective composite.

Further, the data show that once a description is generated it has consequences for subsequent composite production. We included the description delay condition with the expectation that there would be a release from verbal overshadowing, with better performance in the description delay than description no-delay condition. This was not the case. Instead, providing a description of the target face, and then waiting 30 minutes until face construction led to a *reduction* in correct naming. This situation reflects that in Experiment 1 when using a feature-based system, and suggests that once a description has been generated, conditions which interfere with optimal access to the description in memory (i.e., here, a delay to construction) reduce the effectiveness of the composites produced. In fact, the description delay condition compared to both the no description and description no-delay conditions, produced composites that prompted naming less frequently overall (i.e., less frequent correct and mistaken names). This suggests that the composites produced were less likely to resemble any specific identity and so were less likely to attract any name.

Taken together, these findings are consistent with the notion that eliciting a description allows a person to construct a more effective composite. Moreover, once a description of the target has been generated, witnesses will tend to rely upon that description during face construction. Surprisingly, the data show this to be the case even when composites were derived using a holistic system, a system where a description is not integral to the process of producing a face.

GENERAL DISCUSSION

Previous research has indicated that featural descriptions can interfere with a person's ability to recognise a face. Instead, our results imply a beneficial role for descriptions in terms of facilitating successful completion of a different type of face retrieval task: reproduction of a single face from memory. Experiment 1 employed a feature-based method to create a face and showed that, under some circumstances, inserting a 30-minute delay between describing a target face and producing a composite reduced the effectiveness of the likeness produced. In particular, this detriment was evident following intentional compared to incidental encoding (cf. likeness ratings) and following a longer delay after viewing the target face (2 days vs. 3-4 hours; cf. correct naming scores). Experiment 2 found that the descriptions given by participants who constructed composites in Experiment 1 were more effective (i.e., more readily matched to their target photographs by independent judges seeing the descriptions alone) when they were generated following intentional (cf. incidental) encoding, indicating these descriptions contained information more useful for discrimination. However, descriptions were found to be less effective following a shorter (3-4 hour) compared to longer (2 days) post-encoding delay. This result could be considered as surprising, as discussed later in this section, given that we might expect a stronger memory for the face at the short (vs. long) post-encoding delay. When considering these findings with those of Experiment 1, a picture emerges whereby inserting a delay (in this case 30 minutes in duration) between description and composite construction reduces composite effectiveness for those conditions where descriptions proved *more* useful for discrimination (i.e., the intentional encoding conditions and 2 day post-encoding delay conditions).

Experiment 3 involved a holistic-based method for participants to construct a face 22-26 hours after target encoding. This method allowed for the inclusion of a no-description condition and here we expected to observe effects of verbal overshadowing given that holistic (cf. feature) systems place potentially greater emphasis upon recognising faces as whole entities. However, we observed equivalent correct naming rates for both the no description and description no-delay condition. Instead, replicating the findings arising from use of a feature system in Experiment 1, inserting a delay (vs. no-delay) after the description resulted in composites that elicited fewer correct names overall.

The experiments thus found a consistent detrimental influence of inserting a (30-minute) delay between description and face construction using two different pools of target faces, characters from EastEnders (Experiment 1) and retail staff (Experiment 3), and so this surprising result is not tied to a specific set of targets. Frowd and Fields (2011), however, found the opposite, using international footballers as targets. The potential pool of both EastEnders and retail staff targets would have been smaller than for international footballers, as well as being less visually homogenous, and so may have contributed to facilitating naming in the current study. To assess this possibility, we compared naming rates of studies that have used different target pools. Three such studies utilised a similar version of EvoFIT and similar conditions to Experiment 3 (interviewing and composite construction taking place one day after target encoding). Although using different target pools (international footballers, Frowd et al., 2012a; retail staff, Frowd et al., 2012b; EastEnders characters, Frowd et al., 2013), correct naming rates are consistent (ranging from 22.7 - 24.1%) as well as being very similar to that obtained in Experiment 3 (22.5%). Consistency across studies indicates that target pool size does not appear to exert a notable impact on composite naming.

Frowd and Fields (2011) previously identified an effect of verbal overshadowing upon the effectiveness of feature-based composites when applying similar conditions to those investigated here, within Experiment 1: intentional encoding and a 2 day post-encoding delay. However, their effect was small-to-medium in size and only detectable using more sensitive (non-spontaneous) naming tasks, and so unlikely to be of any real forensic value. We did not find any evidence of verbal overshadowing within our data, where target encoding was longer and potentially more effective (including the use of videos compared to photographs). In comparison, their results would seem to apply to the situation where encoding is very brief and the retention interval long: a situation where a particularly weak memory is likely to have formed. Visual memory for the target was already weak, and it appears that under these circumstances relying on verbal recall (that itself is likely to be very limited) was an even less effective strategy when attempting to construct an identifiable target face (i.e., verbal overshadowing was apparent).

Some researchers have proposed that the act of describing a face can lead participants to temporarily alter the way in which they process the face when encountering it again later. Participants apply featural processing at the expense of holistic processing better suited to tasks involving face recognition (e.g., Brown & Lloyd-Jones, 2003; Schooler, 2002). Whilst the process of making a feature-based composite greatly relies upon selecting individual facial features (Frowd et al., 2005), holistic processing is also important (e.g., Frowd et al., 2008). For example, constructors have been shown to produce better-recognised composites if, following their featural description of the target, they also think about the whole face by making personality judgements. Moreover, this whole-face technique has been found to improve the effectiveness of both feature-based and holistic-based composite systems (for a review, see Frowd et al., 2015; note, a subset of the current dataset described in Footnote 4 also supports a benefit of this whole-face technique when using a holistic system). Nevertheless, our findings do not fit well with an account relying upon alterations in processing (e.g., Schooler, 2002). A delay between description and composite construction should have allowed a temporary “switch” to featural processing to subside. Given that holistic processing would have resumed, we should have observed an improvement, not a detriment, in composite effectiveness compared to the description no-delay condition.

Taken together, our findings instead imply a role for the contents of the facial description when determining the effectiveness of the composite produced. The description delay conditions that showed a reduction in composite effectiveness in Experiment 1 were also those conditions that had elicited more useful (i.e., more discriminating) descriptions as demonstrated in Experiment 2. Specifically, these descriptions contained a greater amount of accurate information. Previous research in the face recognition domain has also identified a positive, albeit weak, relationship between the accuracy of the contents of participants’ descriptions and the accuracy of choosing the corresponding face from a line-up task (for discussion, see meta-analysis by Meissner, Sporer, & Susa, 2008). What our findings further imply is that any benefit conferred by relying upon a description in memory may be lost if composite construction does not then take place immediately.

The benefit of providing a description close in time to composite construction (compared to after a 30-minute delay) was apparent for only some conditions. It seems that the timing of the description is not critical on occasions where composite construction occurs on the same day as viewing a crime (approximated by our 3-4 hour post-encoding delay conditions). Typically, more accurate face identification decisions are associated with shorter compared to longer retention intervals (e.g., Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Shapiro & Penrod, 1986), likely to be indicative of a more robust visual representation. Thus, whilst the effectiveness of providing a description for composite construction may decline over the course of a 30-minute delay, it may be that when the retention interval after target encoding is short, the visual memory of the target is strong enough to compensate.

Both visual and verbal memory representations for the face are likely to have subsided at longer delays. When construction takes place 2 days after target encoding, interfering with access to a previously given description (via a 30-minute delay) appears to hamper composite effectiveness. This is perhaps because descriptions given 2 days compared to 3-4 hours after target encoding were more useful for subsequently distinguishing the target face (as found in Experiment 2). This idea is consistent with other research subsequently published whilst the current work was underway. Wilson, Seale-Carlisle and Mickes (2018) found that manipulating retention interval between target encoding and description (no delay vs. 20 minutes delay) led to descriptions differing in their usefulness (as assessed by the success with which independent judges identified the correct target based on a description alone). As here, better recognition performance was associated with those manipulations giving rise to a more diagnostic description.

Whilst Wilson et al. (2018) found descriptions given 20 minutes compared to immediately after target encoding were less useful, Ellis et al. (1980) showed description usefulness remained relatively stable over the course of a day (one hour vs. one day following target encoding). Our data add to this seemingly complex picture and further highlight the variation in how face recall deteriorates over time. The data show that 2 days after target encoding, descriptions can contain information usefully diagnostic of the target face; in this case, more so than descriptions given on the same day as target encoding. Critically, whilst participants recalled less information, this information appears to be more accurate.

We can speculate upon why participants generated fewer incorrect descriptors following a 2 day compared to 3-4 hour post-encoding delay. Research has shown that individuals regulate their recall responses by distinguishing between details they are more or less confident about (Luna & Martín-Luengo, 2012), and information reported with higher confidence tends to be more likely correct (Wixted, Mickes, & Fisher, 2018). Work by

Ebbesen and Rienick (1998) found that the likelihood of person descriptors given with absolute certainty being accurate (compared to those expressed with less confidence) increased as retention interval increased from 1 day to 7 days to 28 days after target encoding. Thus, if we assume witnesses refrain from reporting information about which they are less confident, and less confident memories are more likely to be incorrect, then it seems reasonable to suggest that with the passing of time participants may have reported fewer incorrect descriptors—the pattern reflected within our current dataset. Critically, participants constructed a worse quality composite when access to that more accurate description in memory was hindered (i.e., by a 30-minutes delay). This implies the accuracy of description content is useful to composite construction. Previous research has similarly found an association between fewer incorrect details recalled about the face and better performance on face identification tasks (cf. Meissner et al., 2008).

Whilst a description may serve to augment a decay in the visual memory trace over time, data from the likeness-rating task indicate that low memory strength is not the only factor responsible for determining how verbal recall impacts upon composite construction. Likeness ratings (but not the naming task) showed a description *benefit* under intentional compared to incidental encoding conditions, a condition displaying greater memory strength for the target (here, participants recalled a higher total amount of information about the face). It is possible that the conditions under which participants encode the target affect the success with which the resulting composite accurately portrays featural or holistic information about the face. Participants when intentionally encoding a target tend to report remembering the face by attending to its individual features (Laughery, Duval, & Wogalter, 1986). In contrast, incidental encoding elicits reports of attending to the face as a whole (Olsson & Juslin, 1990). In keeping with this, we found descriptions elicited under intentional (cf. incidental) encoding conditions were not only more diagnostic of identity, but contained more correct featural descriptors. Compared to naming tasks, likeness-rating tasks emphasise a comparison of the similarity of individual facial features between the composite and target photograph (Frowd et al., 2005). On this basis, it seems that intentional (cf. incidental) encoding led to composites containing more accurate feature information, the type of information towards which the likeness-rating task is sensitive. Taken together, our results suggest that verbal recall may be used flexibly to augment visual information about the face, perhaps by supplementing decaying memory traces over time or by directing attention to distinguishing aspects of the visual representation in memory.

Feature-based composite systems are currently used in Europe (E-FIT and PRO-fit), and in the U.S. (FACES and Identikit 2000). For these systems a description is necessary in order to narrow down the visual examples (i.e., noses, eyes, mouth) within the system to which the witness is exposed. The data hint at another benefit for eliciting a description: access to verbally-describable information relating to the offender's face helps the witness to more effectively produce a likeness. Many police forces within the UK use holistic-based composite systems, and thus practically it was important to establish whether describing a perpetrator had similar effects for composites produced using this latter type of system. Specifically, for this type of system, we modelled the situation in which a witness was invited to construct a composite of a target he or she had encoded one day earlier; a typical witnessing experience. We found no evidence of a detriment of providing a description compared to no-description immediately prior to composite construction (i.e., no evidence of verbal overshadowing). However, as for a feature-based system, we found that when a person had described the target face (as would be the case in standard forensic practice), then delaying composite construction (in this case by 30-minutes) resulted in the production of less effective composites.

The processes utilised for producing feature-based and holistic-based composites on the surface appear very different. However, there is evidence arising from both systems, which indicate that when feature information is better utilised, composite effectiveness improves. For example, the effectiveness of composites produced using holistic systems, like feature systems, can improve by initially encoding target faces in terms of their features rather than via whole face judgements (Frowd et al., 2007b). This suggests that access to feature information in memory can similarly be useful for producing composites using either method. Our findings concerning the utility of describing faces for composites constructed using both types of police system support this assumption. More specifically, they suggest that inserting an appreciable delay (here, 30 minutes) impairs access to details of recalled descriptions that are valuable for face construction with both methods of production. In addition, research by Frowd et al. (2012b) indicates that asking constructors to make personality judgments about their target face prior to holistic face construction (as mentioned above) is only effective (to produce composites with higher correct naming) when preceded by a free description of the face. Thus, their experiment again demonstrates the importance of a facial description as, without recalling it, a constructor’s composite is not more effective. More generally, these findings perhaps suggest that refreshing the witness’s memory for their description of the face may be a fruitful intervention for overcoming detrimental effects of description delay upon composite effectiveness. A straightforward way to do this would be to request an additional free recall from a witness immediately prior to construction of a facial composite. Such a technique may bring valuable details to mind and allow effective use of this information during face construction; research on this issue is in progress.

Regardless of whether a feature or holistic-based system is utilised, it is standard practice for police practitioners to elicit a description of the person of interest from the witness prior to constructing a composite. Whilst current police composite procedures do not explicitly recommend giving the witness a break between providing a description and composite construction, there are many situations where police officers simply offer this option. For example, witnesses can find the process of providing a description of the offender effortful or stressful. Intuitively, providing a break to the witness before beginning the long and mentally demanding procedure of building a composite may provide some relief. However, our findings clearly show that inserting a delay can be detrimental under situations common within forensic settings: When constructed one or two days following the crime using either a holistic- or feature-based composite system. In these instances describing the face and then delaying composite construction (here, by 30 minutes) more than halves the likelihood that the constructed composites will be successfully named (*Exp(B)* > 2). In a forensic situation, this result translates into a sizeable reduction in correct names given to composites, with the knock-on effect of reducing accurate leads in a criminal investigation. Our findings therefore show that access in memory to a description of the perpetrator *does* matter under these circumstances. Thus, the practical message arising from our work is that witnesses should produce their composite immediately, without appreciable delay, after recalling the appearance of the offender.

Acknowledgements

The UK Economic and Social Research Council provided support for some aspects of this work [ES/I002022/1] and the dataset supporting the conclusions of Experiment 1 of this article is available in the UK Data Service repository, http://doi.org/10.5255/UKDA-SN-850883. Full data sets are available on request from the authors.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*(5), 556–578.

Brace, N., Pike, G., & Kemp, R. (2000). Investigating E-FIT using famous faces, in Czerederecka, A., Jaskiewicz-Obydzinska, T., & Wojcikiewicz, J. (Eds), *Forensic Psychology and Law*, Krakow, Institute of Forensic Research Publishers, pp. 272–6.

Ballinger, G.A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods, 7 (2),* 127–150.

Brown, C., & Lloyd-Jones, T. J. (2003). Verbal overshadowing of multiple face and car recognition: Effects of within-versus across-category verbal descriptions. *Applied Cognitive Psychology, 17*(2), 183–201.

Brown, C., Lloyd-Jones, T. J., & Robinson, M. (2008). Eliciting person descriptions from eyewitnesses: A survey of police perceptions of eyewitness performance and reported use of interview techniques. *European Journal of Cognitive Psychology*, *20*(3), 529–560.

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.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cui, J., & Qian, G. (2007). Selection of working correlation structure and best model in GEE analyses of longitudinal data. *Communications in Statistics - Simulation and Computation, 36*(5), 987–996.

Davies, G.M, & Hine, S. (2007). Change blindness and eyewitness testimony. *The Journal of Psychology, 141*(4), 423–434.

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.

Deffenbacher, K.A., Bornstein, B.H., McGorty, E.K., & Penrod, S. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied, 14*(2), 139–150.

Diamond R., & Carey S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107–117.

Ebbesen, E.B., & Rienick, C.B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology, 83 (5),* 745–762.

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.

Ellis, H.D., Shepherd, J.W., & Davies, G.M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. *Journal of Police Science and Administration, 8*(1), 101–106.

Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1608–1623.

Finger, K., & Pezdek, K. (1999). The effect of cognitive interview on face identification accuracy: Release from verbal overshadowing. *Journal of Applied Psychology, 84*(3), 340–348.

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., Bruce, V., McIntyre, A., & Hancock, P.J.B. (2007a). The relative importance of external and internal features of facial composites. *British Journal of Psychology, 98*(1), 61–77.

Frowd, C.D., Bruce, V., Ness, H., Bowie, L., Paterson, J., Thomson-Bogner, C., McIntyre, A., & Hancock, P.J.B. (2007b). Parallel approaches to composite production: interfaces that behave contrary to expectation. *Ergonomics, 50* (4), 562–585.

Frowd, C.D., Bruce, V., Smith, A.J., & Hancock, P.J.B. (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., Richardson, J., Morrison, L., McLanaghan, S., & Hancock, P.J.B. (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.B. (2015). A decade of evolving composites: regression-and meta-analysis. *Journal of Forensic Practice*, *17*(4), 319–334.

Frowd, C. D., & Fields, S. (2011). Verbalization effects in facial composite production. *Psychology, Crime & Law, 17*(8), 731–744.

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., 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., Skelton, F., Atherton, C., Pitchford, M., Hepton, G., Holden, L., McIntyre, A., Hancock, P.J.B., (2012a). Recovering faces from memory: the distracting influence of external facial features, *Journal of Experimental Psychology*. *Applied, 18,* 224–238.

Frowd, C.D., Nelson, L., Skelton, F.C., Noyce, R.. Atkins, R., Heard, P., Morgan, D., Fields, S., Henry, J., McIntyre, A., Hancock, P.J.B. (2012b). Interviewing techniques for Darwinian facial composite systems, *Applied Cognitive Psychology, 26*(4), 576–584.

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.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1986). Eyewitness memory enhancement with the cognitive interview. *American Journal of Psychology, 99,* 385–401.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517.

Laughery, K.R., Duval, C., & Wogalter, M.S. (1986). Dynamics of facial recall. In *Aspects of Face Processing*. H.D. Ellis, M.A. Jeeves, F. Newcombe and A. Young (Eds.), pp. 373–387. Dordrecht: Martinus Nijhoff.

Luna, K., & Martín-Luengo, B. (2012). Confidence–accuracy calibration with general knowledge and eyewitness memory cued recall questions. *Applied Cognitive Psychology, 26*, 289–295.

Malpass, R. S., Lavigueur, H., & Weldon, D. E. (1973). Verbal and visual training in face recognition. *Perception & Psychophysics, 14*(2), 285–292.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*(6), 603–616.

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition, 29*(1), 176–186.

Meissner, C.A., Sporer, S. L., & Susa, K.J. (2008). A theoretical review and meta-analysis of the description-identification relationship in memory for faces. *The European Journal of Cognitive Psychology, 20*(3), 414–455.

Ness, H., Hancock, P. J., Bowie, L., Bruce, V., & Pike, G. (2015). Are two views better than one? Investigating three-quarter view facial composites. *Journal of Forensic Practice, 17*(4), 291–306.

Olsson, N., & Juslin, P. (1999). Can self-reported encoding strategy and recognition skill be diagnostic of performance in eyewitness identifications? *Journal of Applied Psychology, 84*(1), 42–49.

Osborne, J.W. (2006). Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses. *Practical Assessment Research & Evaluation, 11(7).* Available online: http://pareonline.net/getvn.asp?v=11&n=7.

Pan, W. (2001). Akaikes Information Criterion in Generalized Estimating Equations. *Biometrics, 57*(1), 120–125.

Penrod, S. D., Loftus, E. F., & Winkler, J. (1982). The reliability of eyewitness testimony: A psychological perspective. In N. Kerr & R. Bray (Eds.), *The Psychology of the Courtroom* (pp. 119–168). New York: Academic Press.

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology, 16*(8), 989–997.

Schooler, J. W. (2014). Turning the lens of science on itself: Verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science, 9*(5), 579–584.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*(1), 36–71.

Shapiro, P. N., & Penrod, S. D. (1986). Meta-analysis of facial identification rates. *Psychological Bulletin, 100*(2), 139–156.

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 (pp. 3 – 37). Psychology Press: New York.

Tanaka, J.W., & Farah, M.J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, A46*, 225–245.

Tanaka, J.W., & Sengco, J.A. (1997). Features and their configuration in face recognition. *Memory & Cognition, 25*, 583–592.

Tredoux, C. G., Nunez, D. T., Oxtoby, O., & Prag, B. (2006). An evaluation of ID: an eigenface based construction system. *South African Computer Journal*, *37,* 1–9.

Turtle, J. W., & Yuille, J. C. (1994). Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hypermnesia. *Journal of Applied Psychology*, *79*(2), 260–271.

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.

van Koppen, P. J., & Lochun, S. K. (1997). Portraying perpetrators: The validity of offender descriptions by witnesses. *Law and Human Behavior*, *21*(6), 661–685.

Wells. G.L., & Hryciw, B. (1984). Memory for faces: Encoding and retrieval operations. *Memory & Cognition, 12*(4), 338–344.

Wilson, B.M., Seale-Carlisle, T.M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of Experimental Psychology: General, 147(1),* 113–124.

Wixted, J.T., Mickes, L., & Fisher, R.P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science, 13(3),* 324–335.

Yarmey, A. D. (2004). Eyewitness recall and photo identification: A field experiment. *Psychology, Crime and Law*, *10*(1), 53–68.

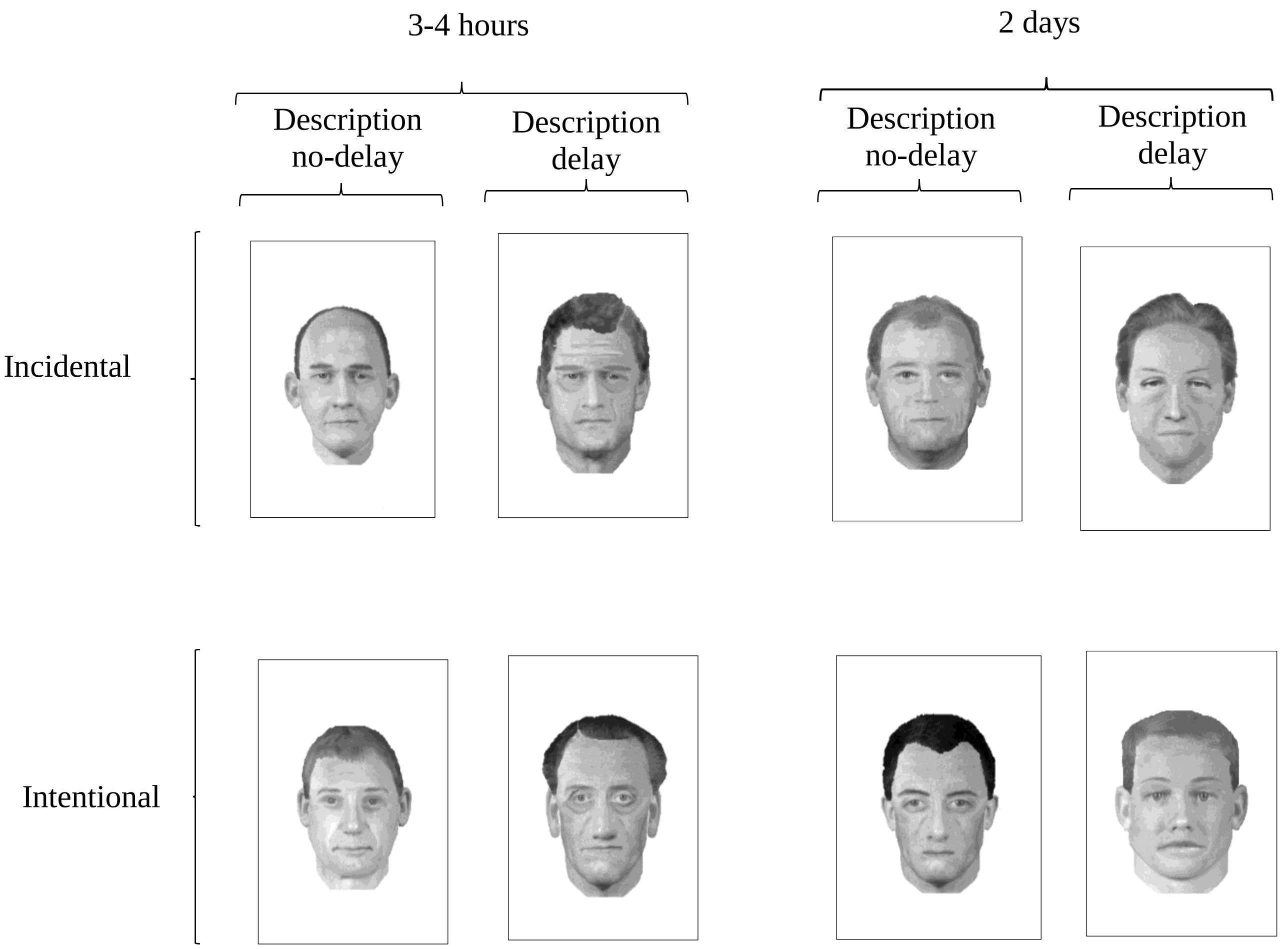

Figure 1. Composites constructed to resemble 'Billy Mitchell' from the BBC TV programme 'EastEnders'. Composites were constructed: (i) by encoding condition (intentional vs. incidental), (ii) post-encoding delay (3-4 hours vs. 2 days) and (iii) interview (description no-delay vs. description delay). These composites (along with other composites produced in the study) were given to fans of EastEnders to name.

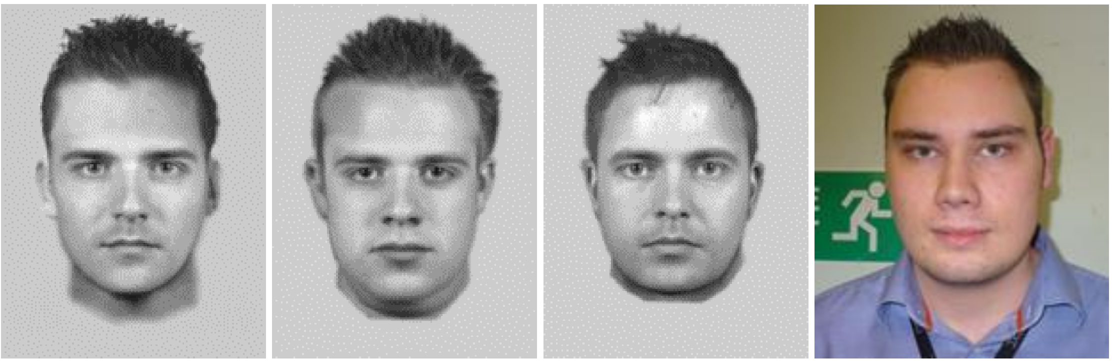

Figure 2. Example composites constructed from a holistic system of one of the targets (far right) used in the experiment. Each composite was produced by a different participant 22-26 hours after having watched a video of the target person giving directions to a local town centre. From left to right, the composite was produced without providing a description of the face (no-description), after freely recalling the face (description no-delay) and 30 minutes after free recall (description delay).

Table 1: Percentage correct naming for composites constructed from a feature system (Experiment 1) as a function of encoding, post-encoding delay and (delay or not following) interview.

Post-Encoding Delay		*Interview*	
	Encoding	description no-delay	description delay
3-4 hr			
	Incidental	15.6 (15 / 96)	8.9 (8 / 90)
	Intentional	16.7 (16 / 96)	16.5 (14 / 85)
2 days			
	Incidental	13.8 (13 / 94)	3.2 (3 / 93)
	Intentional	14.7 (14 / 95)	5.4 (5 / 93)

Note. Figures are percentage-correct naming calculated from responses in parentheses: summed correct responses (numerator) and total (correct plus incorrect) responses (denominator). These data are for composites for which participants correctly named the associated target (N = 742 out of 768). See text for predictors retained within the final model. Generalised Estimating Equations, intercept [$X^2(1) = 309.26$, $p < .001$]).

Table 2. Percentage correct naming for composites constructed from a feature system (Experiment 1) as a function of post-encoding delay and (delay or not following) interview.

	Interview	
	description no-delay	description delay
Post-Encoding Delay		
3-4 hours	16.1[c] (31 / 192)	12.6[b] (22 / 175)
2 days	14.3[a] (27 / 189)	4.3[a,b,c] (8 / 186)

Note. See Table 1. [a,b,c] $p < .005$.

Table 3: Percentage mistaken naming for composites constructed from a feature system (Experiment 1) as a function of encoding, post-encoding delay and (delay or not following) interview.

Post-Encoding Delay		*Interview*	
	Encoding	description no-delay	description delay
3-4 hr			
	Incidental	46.9 (38 / 81)	46.3 (38 / 82)
	Intentional	40.0 (32 / 80)	46.5 (33 / 71)
2 days			
	Incidental	34.6 (28 / 81)	24.4 (22 / 90)
	Intentional	40.7 (33 / 81)	46.6 (41 / 88)

Note. See Table 1, except here data are for composites for which participants correctly named the associated target photograph, but for which they provided a mistaken name (N = 265) or ‘don’t know’ response (N = 389). Details of predictors included in the final model, Generalised Estimating Equations: Encoding [$X^2(1) = 0.94$, $p = .33$], Post-encoding delay [$X^2(1) = 2.00$, $p < .16$], Interview [$X^2(1) < 0.01$ $p < .98$] and Intercept [$X^2(1) = 9.75$, $p < .002$].

Table 4: Ratings of likeness between composites constructed from a feature system (Experiment 1) and corresponding target photographs as a function of encoding, post-encoding delay and (delay or not following) interview.

Post-Encoding Delay		*Interview*	
	Encoding	description no-delay	description delay
3-4 hr			
	Incidental	3.36 (1.55)	3.31 (1.53)
	Intentional	3.46 (1.04)	3.11 (1.16)
2 days			
	Incidental	3.31 (1.75)	3.28 (1.56)
	Intentional	3.60 (1.41)	2.93 (1.27)

Note. Values are mean participant ratings (1 = poor likeness; 10 = good likeness). See main text for details of analysis.

Table 5: Percentage of descriptions (generated by participant-constructors in Stage I of Experiment 1) correctly matched to photographs of the target face (Experiment 2) as a function of encoding, post-encoding delay and the type of recall elicited (free recall or free and cued recall).

Post-Encoding Delay	*Encoding*	*Recall Type* Free Recall	 Free and cued recall
3-4 hr			
	Incidental	41.7 (60 / 144)	41.7 (60 / 144)
	Intentional	53.5 (77 / 144)	51.4 (74 / 144)
2 days			
	Incidental	49.3 (71 / 144)	43.8 (63 / 144)
	Intentional	57.6 (83 / 144)	65.3 (94 / 144)

Note. Figures are percentage correct matches calculated from responses in parentheses: summed correct responses (numerator) and total (correct plus incorrect) responses (denominator). See text for predictors retained in the final model. Generalised Estimating Equations, intercept [$X^2(1) = 0.11$, $p = .74$].

Table 6: Percentage correct naming for composites constructed from a holistic system as a function of type and timing of the face-recall interview

no-description	description no-delay	description delay
21.7 (26 / 120)	22.5 (27 / 120)	11.7* (14 / 120)

Note. See Table 1. Participants correctly named all target photographs used to construct the composites (*N* = 360 out of 360). Interview was a significant predictor (see details in text), in addition, Generalised Estimating Equations intercept [$X^2(1) = 248.25$, $p < .001$]. *Different from the other two conditions ($p < .005$).

Table 7: Mistaken naming (percentage) for composites constructed from a holistic system as a function of type and timing of face-recall interview

no-description	description no-delay	description delay
80.9*† (76 / 94)	68.8† (64 / 93)	58.5*† (62 / 106)

Note. See Table 6, except here data are for composites for which participants correctly named the associated target identity, but for which they provided a mistaken name (N = 202) or ‘don’t know’ response (N = 91). Interview was a significant predictor (see details in text). Generalised Estimating Equations intercept [$X^2(1) = 36.63$, $p < .001$). *Different to each other ($p = .001$). †Significant linear trend in the order shown ($p < .001$).

FOOTNOTES

1. For all GEE analyses reported, we note the minimal change in QIC values when comparing use of an exchangeable (for related data) and independent (non-related) correlation matrix. To further validate outcomes of GEE, all analyses were repeated using an independent (cf. exchangeable) working correlation matrix. This did not change interpretation of the findings.
2. The Odds Ratio (effect size) reported within the text are calculated by exponentiating the variable's slope parameter *B*, *Exp(B)*. Variables with the lowest numerically coded-category were used as the reference category. Thus, positive or negative values of *B* may arise depending on how variables were coded. Negative values of *B* are accompanied by Odds Ratio Effect Sizes [*Exp(B)*] that are less than 1.0 reflecting a decrease in the odds of experiencing an outcome. Unlike increasing odds ratios which can vary from 1.0 to infinity, decreasing odds ratios are restricted ranging from 0 to 1. Therefore, to standardise interpretation of effect sizes throughout, where *Exp(B)* < 1, we take the inverse (expressed as 1/*Exp(B)*) which provides a corresponding ratio greater than 1 (cf., Osborne, 2006).
3. We asked eight separate participants to freely describe each target from a video still and to rate each face on a series of five-point bimodal scales (e.g., short–long, dark–light) relating to 10 separate features (overall face shape, complexion, hair, forehead, eyebrows, eyes, nose, mouth, ears and chin). Details freely and consistently mentioned by four or more participants were classed as correct, as were descriptors consistently rated by five or more participants. Two coders independently coded a subset of 32 face descriptions against this protocol and resolved any discrepancies. One coder went on to code the remaining face descriptions. Inter-rater reliability was

high (*p* < .001): correct details, *r* = .83; incorrect details, *r* = .80; subjective details, *r* = .92; and total number of details, *r* = .97.

4. A fourth condition, holistic recall, was included in Experiment 3, and while not directly relevant to the current work, we report the outcome of the correct naming of these composites here. This condition required participants to rate a target face in terms of personality traits they would attribute (e.g., honest, masculine) immediately after describing the face, but prior to producing a composite. Previous work has found this manipulation to improve correct naming rates to composites produced using a holistic composite system compared to a condition similar to our description no-delay condition (for a review, see Frowd et al., 2015). Our finding regarding this condition is in keeping with this research. We used GEE to fit a binary logistic regression model that included the categorical variable, interview, with two levels: description no-delay and holistic recall. Interview was a significant predictor [$X^2(1) = 7.13$, *p* = .008]. Correct naming scores to composites produced in the holistic recall condition (35.8%) were significantly higher than those obtained in the description no-delay condition (22.5%) [*B* = 0.65, *SE(B)* = 0.25, *Exp(B)* = 1.92 (1.19, 3.11)].