Legal and Criminological Psychology (2005), 10, 63–81

The British Psychological Society

www.bpsjournals.co.uk

Contemporary composite techniques: The impact of a forensically-relevant target delay

Charlie D. Frowd[1]*, Derek Carson[2], Hayley Ness[1], Dawn McQuiston-Surrett[3], Jan Richardson[4], Hayden Baldwin[5] and Peter Hancock[1]

[1]Department of Psychology, University of Stirling, UK
[2]Department of Psychology, University of Abertay, Dundee, UK
[3]Department of Psychology, Arizona State University West, USA
[4]Evidential Art, Market Harborough, Leicestershire, UK
[5]Forensic Enterprises, Orland Park, IL, USA

Purpose. Previous laoratory-based research suggests that facial composites, or pictures of suspected criminals, from UK computerized systems are named correctly about 20% of the time. The current work compares composites from several such systems following a more realistic interval between seeing an 'assailant' and constructing a composite. Included are those used by police in the UK (E-FIT, PROfit and sketch), and the USA (FACES), and a system in development (EvoFIT).

Method. Participant-witnesses inspected a photograph of a celebrity for 1 minute and then 2 days later constructed a composite from one of these systems using a procedure closely matching that found in police work; for example, the use of a Cognitive Interview and computer operators/artists who were appropriately trained and experienced. Evaluation was assessed mainly by asking independent observers to name the composites. Two common auxiliary measures were used, requiring composites to be matched to their targets (sorting), and photographs to be chosen from an array of alternatives (line-up).

Results. Composite naming was surprisingly low (3% overall), with sketches named best at 8%. Whereas composite sorting revealed a broadly similar pattern to naming, photo line-ups gave a poor match.

Conclusion. With a 2 days delay to construction, the results suggest that, while likenesses can be achieved, few composites would be named in police work. The composite sorting data provide further evidence that the computerized systems tested perform equivalently but are poorer than the manually-generated sketches. Lastly, the data suggest that line-ups may be a poor instrument for evaluating facial composites.

*Correspondence should be addressed to Charlie Frowd, Department of Psychology, University of Stirling, FK9 4LA, UK (e-mail: cdf1@stir.ac.uk).

DOI:10.1348/135532504X15358

Witnesses to crime are often faced with the difficult task of describing the appearance of a suspect, typically only observed for a short time. In serious cases such as sexual assault and murder, witnesses are often called forward a number of days or even weeks later to construct a *facial composite*. This is a visual representation of a perpetrator's face and often has a key place in solving a crime, frequently appearing in newspapers and on TV crime programmes. Facial composites also have theoretical interest, giving insight into the psychological processes involved in their construction.

The first widely used technique for constructing a facial composite was a sketch artist. This person would be skilled in portraiture and would draw a composite by hand using pencils or crayons. Other approaches were developed to allow composites to be produced by those less artistic. Two well-known techniques that became popular during the 1970s are Identikit and Photofit. Whereas Identikit was used in the USA and originally contained sketch-like facial features, Photofit was adopted primarily in the UK and contained photographed facial elements.

These two systems have attracted considerable empirical research. The basic methodology used required a participant-witnesses to inspect an unfamiliar target face and then work with a trained technician, or composite operator, to construct a composite. The resulting composites would be evaluated by independent observers, sometimes referred to as composite judges, who (a) assigned a likeness score to the composites in the presence of the original photograph (likeness rating), (b) matched the composites to target photographs (sorting), or (c) selected the composites from among alternatives (photo line-ups). Overall, composite quality was poor and generally independent of manipulations that normally give rise to measurable change elsewhere in face perception (e.g. Christie, Davies, Shepherd, & Ellis, 1981; Davies, Ellis, & Shepherd, 1978; Ellis, Davies, & Shepherd, 1978; Laughery & Fowler, 1980).

Computerization of these kits promoted larger libraries of features and greater flexibility (features could be resized and positioned more easily). One such system, Mac-A-Mug Pro, contained sketch-like features that, unlike Identikit and Photofit, produced high-quality composites when a target was available for reference during construction (Cutler, Stocklein, & Penrod, 1988). However, when participant-witnesses worked from their memory of the target, using procedures similar to the Identikit/Photofit research described above, composite quality was once again poor (Koehn & Fisher, 1997), even if that person was well-known to the participant (Kovera, Penrod, Pappas, & Thill, 1997).

Alternative systems now exist with photographed features, producing more realistic-looking faces. Examples of such systems in use by law enforcement agencies include E-FIT, PROfit, FACES and Identikit 2000. Although varying in ecological validity, recent laboratory-based research has reported that composites from these systems are good enough to be spontaneously named about 20% of the time (e.g. Brace, Pike, & Kemp, 2000; Bruce, Ness, Hancock, Newman, & Rarity, 2002; Davies, van der Willik, & Morrison, 2000; Frowd, Hancock, & Carson, 2004).

Of this recent work, arguably the most realistic design is that of Frowd *et al.* (in press). They compared five composite techniques: E-FIT, PROfit, sketch, Photofit and EvoFIT. Suitably experienced artists and computer operators (one for each composite type) each worked with participant-witnesses to construct a set of composites following UK procedures specified by the Association of Chief Police Officers (Scotland) Working Group [ACPOS] (2000). Participant-witnesses first inspected a photograph of a celebrity that they did not know for 1 minute. After 3 to 4 hours, operators administered a Cognitive Interview, used to assist the recall of accurate information (e.g. Geiselman, Fisher, MacKinnon, & Holland, 1986). Operators then worked with witnesses to construct the best possible likeness of the original target face. Fifty composites were constructed (10 from each system) and these were given to participant-judges to name. They reported the best naming for E-FIT and PROfit (composites were correctly named about 20% of the time), followed by artist's sketches (10%), Photofit (5%) and EvoFIT (3.5%). A sorting task was also employed, requiring further participants to match the composites to their target photographs, and yielded a similar pattern of results.

Despite the level of realism, this research only utilized a 3–4 hour delay between encoding and interview—previous studies were even shorter. The current work employed a 2-day delay, which is more typically found in criminal investigations. Given a longer delay, one might suppose that composite quality would suffer due to a decline in participants' memory; certainly composites from Mac-A-Mug Pro were very poor after such an interval (Koehn & Fisher, 1997). Curiously, composite quality from Photofit and Identikit was not found to deteriorate even after a week or more (Davies *et al.*, 1978; Green & Geiselman, 1989; McNeil *et al.*, 1987), although this is arguably due to the poor renditions of these systems.

The general finding in face recognition does suggest performance decrements with increasing delay (e.g. Krouse, 1981; Shepherd, Gibling, & Ellis, 1991), and that this decay is linear (Shapiro & Penrod, 1986). However, such deficits tend not to be measurable for several weeks or even months (e.g. Laughery, Fessler, Lenorovitz, & Yoblick, 1974; Shepherd, 1983). In contrast, Ellis, Shepherd, and Davies (1980) found that participants recalled noticeably less detail after only 24 hours. Similarly, participants undertaking a Cognitive Interview, used to enhance a witness's memory of a suspect, have also been found to recall less overall information about a target face after 24 hours (Carson, 2000). Given that composite construction involves both recall (initially describing a face; later verbalizing changes to improve the composite) and recognition (selection of facial features)—and that recall is likely to be reduced—composites from the current study (2 days delay) are likely to be worse than those constructed after a much shorter delay (3–4 hours or less).

The current study adopted a methodology that, as far as possible, reflected composite construction in the real world. For example, the work utilized experienced composite operators and a Cognitive Interview. Five systems were included—E-FIT, PROfit, FACES, sketch and EvoFIT—given the availability of suitably trained and experienced operators (see below). E-FIT and PROfit are very similar systems and produce composites of equivalent quality (e.g. Bruce *et al.*, 2002; Davies *et al.*, 2000; Frowd *et al.*, 2004). With these systems, witnesses describe an assailant's face, via a Cognitive Interview, and then select facial features to match. Both systems contain a large number of photographed features for hair, face shape, eyes, brows, nose, mouth and ears. These features are always selected in the context of a whole face, rather than as a set of individual features, to encourage natural *holistic* face processing, (e.g. Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Also, selected features may be resized and positioned as required. Lastly, both systems make use of software drawing packages to improve the likeness of the face, such as the addition of stubble.

To construct a sketch, witnesses also describe the face using a Cognitive Interview. They then select appropriate feature shapes and work on the proportions of a face. Features are next drawn out by hand, initially as an outline drawing and then in more detail. In general, this approach develops detail for groups of features, rather than individual features. As such, sketching has been found to produce better quality composites when participant-witnesses have observed a target more as a whole face rather than concentrating on individual features (Davies & Little, 1990), a holistic type of face processing.

EvoFIT was designed to capitalize on our natural holistic face processing. This system, currently under development at the University of Stirling, contains a model built from faces in their entirety using a statistical technique called principal components analysis (e.g. Frowd, 2002, Frowd, Hancock & Carson, 2004; Hancock, 2000, Hancock & Frowd, 2002). The system is based on the tenet that it is hard to describe a face and select individual features, but relatively easy to recognize faces that look similar to one seen previously. To construct an EvoFIT composite, witnesses first select a hairstyle (the model is poor at representing hair) and then view about 70 faces with random characteristics. Witnesses choose about six that look most like the assailant and EvoFIT 'breeds' these faces together to produce another similar sized 'population.' Witnesses choose from this set, and these are then bred together until an acceptable likeness emerges. A composite is thus created by 'evolution' (hence Evolution-FIT, or EvoFIT). In practice, to accelerate the process, witnesses first select facial shapes, then facial intensities or textures; witnesses also choose one face to be the best likeness from these selected shapes and textures: greater priority is then given to this chosen face during breeding. Software tools are also available to manipulate features on demand, for example to make a face longer or rounder.

Unfortunately, despite its appeal, EvoFIT performed rather poorly in Frowd *et al.* (2004). It turned out that despite operator training, the software was rather complex to use. This was especially relevant to the Feature Shift tool used to manipulate features. This utility allows a witness to make a face wider or to move the eyes together, for example, and thereby accelerate evolution and the construction of a composite. Since then, improvements have been made and the software is now more ergonomic.

The fifth technique included in the study was FACES (version 3.0), a computerized system similar to E-FIT and PROfit though much less expensive (about $50 — the other computerized systems cost thousands of pounds). The FACES system also contains photographic-quality features, but unlike PROfit and E-FIT, employs one large mixed gender, age and race feature library — E-FIT and PROfit have separate libraries for gender, age and race. In addition, although features can be resized and repositioned, these functions are less flexible than PROfit and E-FIT. Further, the brightness and contrast of features cannot be altered. It is simple to use and IQ Biometrix, who market the product, claim that no special training for composite operators is necessary (FACES, 2003). The system is clearly popular, with over 150,000 copies sold worldwide. In a recent survey, McQuiston and Malpass (2000) found that FACES had a high reported use in the USA (second only to Identikit 2000).

In summary, the current study evaluated five facial composite techniques under forensically relevant conditions, including a 2 days delay-to-interview. With a reduction in ability to recall details about a target face, we expected poorer quality composites for E-FIT, PROfit and sketch than that found elsewhere (i.e. less than 20% for E-FIT and PROfit; less than 10% for sketch). With operational problems resolved, EvoFIT was expected to be more like E-FIT and PROfit. Being a relatively simple product, FACES was likely to be the worst.

Composite construction

Method

Participant-witnesses were shown a photograph of a face, then 2 days later described this face via a Cognitive Interview and constructed a composite using one of the five systems. The same set of 10 photographs was used as the target for each system, and therefore 10 participant-witnesses constructed composites for each system (50 participants in total). As different participants featured throughout, the study was a between-subjects design for construction technique (E-FIT/PROfit/sketch/EvoFIT/FACES). While this investigation was laboratory-based, the design adhered as closely as possible to the procedures used in a real investigation, thus increasing ecological validity.

Famous faces were chosen as targets, allowing an important measure of composite quality: un-cued or spontaneous naming, the acid test of a composite system. As witnesses who construct composites are generally unfamiliar with assailants, our participant-witnesses were pre-screened to ensure that they were unfamiliar with at least one target. Also, as composite quality appears to improve with practice (Wogalter & Marwitz, 1991), participant-witnesses were selected who had not constructed a composite previously—the norm for real witnesses.

In the UK, care is taken to engage different operators if the same assailant is suspected. This limits the possibility of operators being influenced by previously constructed composites of the same suspect—sometimes referred to as operator contamination. Similarly, different operators were employed for each system in the current design, and therefore operators only constructed one composite per target. Also, given research suggesting that large differences in operator experience can influence the quality of composites (e.g. Davies, Milne, & Shepherd, 1983), our operators attended an accredited course and are experienced (they are also authors of this paper).

In more detail, the sketch artist completed a recognized course in facial composites in the late 1980s and several courses thereafter. She is a very experienced artist, having been constructing composites in the UK with real witnesses for about 15 years, and has ensured that our procedure follows those specified by the ACPOS (2000). The E-FIT and PROfit operators are laboratory-based, have attended a UK accredited training course, and have constructed composites for several years thereafter, many in the context of empirical studies such as the present study. The EvoFIT operator in the current work differed from the one used by Frowd *et al.* (2004). He is also laboratory-based with several years experience of composite construction and operational procedures. Note that as EvoFIT is a developmental system, there is no accredited training course available. The FACES operator is very experienced, having been certified in the late 1980s, and involved with the construction of composites in police work since then. All operators were provided with a written brief of our procedure.

Selection of target photographs

Ten target photographs were used, similar in number to other composite research (e.g. Cutler *et al.*, 1988; Frowd *et al.*, 2004; Kovera *et al.*, 1997). It is well documented that salient or distinctive targets are better remembered than more average-looking faces (Hancock, Burton, & Bruce, 1996; Shapiro & Penrod, 1986; Valentine & Bruce, 1986; Valentine & Endo, 1992). This notion appears to extend to modern composite systems, as composites of distinctive faces are better recognized than composites of average faces (Frowd *et al.*, 2004), though not to an early version of Identikit (Green & Geiselman, 1989).

We identified 18 famous face targets (relatively well-known to our undergraduate students) and collected distinctiveness ratings from 18 undergraduates (1 = *average, blends in to the crowd* and 7 = *very distinctive, stands out from the crowd*).

As familiarity of a face influences distinctiveness judgements (Vokey & Read, 1992), after collecting distinctiveness scores from each person, we also collected familiarity (of the target) scores using a 5-point familiarity rating scale (1 = *Face not known at all*; 2 = *Seen face only [don't know anything about this person]*; 3 = *Could say something about why he is famous, but not very confidently*; 4 = *Definitely known, could say who he is and why he is famous with confidence*; 5 = *Lots known about him and the reasons for his fame*). Then, distinctiveness scores were only analysed from participants who were not very familiar with targets (a familiarity rating of 3 or less). Ten targets were extracted with a range of mean distinctiveness from 2.7 to 4.3 ($SD = 1.4$). The chosen celebrities were actors (Ben Affleck, Matt Damon, Jeremy Edwards, Joshua Jackson, Philip Olivier and James Redmond) and singers in pop groups (Ronan Keating, Kian Egan, Mark Feehily and Ian 'H' Watkins). At the time of construction, this set had an average age of 26.3 years ($SD = 3.9$). Five sets of the pictures, which were largely frontal view and neutral expression, were printed in colour on the same high-quality printer.

Participants

Fifty participants who had not constructed a composite previously were witnesses in the study. There were 17 males and 33 females and their age ranged from 17 to 77 years ($M = 37.8$, $SD = 13.7$). Each person was paid £10 for their time.

Procedure

Participants were informed that the study used famous faces to allow composites to be named by third persons, as in real life. The study was also quite realistic, without stressing witnesses, and so the celebrity faces must be unfamiliar to witnesses. Therefore, the first stage would be to identify an unfamiliar famous face; they would then be given a good opportunity to study the face. Also, participants were requested not to reveal the identity of any famous face, thus allowing operators to be blind to targets.

Participants were shown the target photographs sequentially until they located the first unfamiliar face. Once an unfamiliar face was identified no further target photographs were presented. If all targets were known, the person was thanked and dismissed. Otherwise, the participant-witness inspected the unknown celebrity for 1 minute. This was carried out without the operator knowing the identity of the target, and all operators remained blind until all the composites had been constructed. No guidance was offered to the participant concerning the rehearsal of target faces during the delay interval, as would be the case in real life.

Two days later, participant-witnesses met up with their respective operator. Operators gave an overview of the Cognitive Interview (CI), informing them that the procedures used would follow the normal two-part interview used in a real investigation: a CI followed by composite construction. It was explained that shortly witnesses would be asked to think back to the time when their target was seen and form a mental picture of his face, known as *reinstatement of context*. They were informed that when a clear picture had been formed, they would be asked to describe the target's face, in their own time and with minimal interference from operators, a process called *free recall*. They were made aware that operators would be making notes during recall. Witnesses were requested not to speak too quickly, as this may interfere with recall and make note-taking difficult. This phase would be followed by *cued recall*, whereby descriptions for each feature would be read aloud and participant-witnesses asked whether anything further could be remembered. Once this descriptive phase was completed, the session would then move on to composite construction.

After any questions had been answered, the CI was carried out. Participant-witnesses were encouraged to think back to the time when their target was shown (reinstatement of context) and form a mental image of his face. When formed, they were asked to describe the target's face (free recall) while operators made appropriate notes using the witnesses' own words. When participant-witnesses stated that they could not remember anything further, they were asked to repeat the exercise, as multiple attempts allow more information to be recalled. The operators wrote down any extra information that was not mentioned during the first recall attempt. Cued recall was then carried out, with operators reading aloud descriptions for each feature and asking whether anything further could be recalled. After the witnesses had exhausted further recall for a feature, operators would then make requests if general information had been omitted. This part used open-style questions to probe for shape and colour information, as appropriate. For example, 'OK, you recalled that his eyes were almond-shaped and slightly wide-apart, what do you remember about the colour of the eyes?' or, 'You said that the mouth was small, but in what way was it small?' Once again, operators wrote down any extra information recalled. In general, the CI took approximately 20 minutes to administer.

The session moved on to composite construction, with artist/operators providing an overview of the relevant construction technique. For the contemporary computerized systems (E-FIT, PROfit and FACES) operators also gave a short demonstration of the software, illustrating how features could be selected, re-sized and repositioned. It was explained that the facial features available were not taken from the target's face and so an exact copy would not be possible. However, as there were many examples of each feature, witnesses should work towards the best possible likeness. If a given feature was not available, they should select the best, as a paint package was available to improve the likeness. Participant-witnesses were encouraged to select and position all facial features before any artistic enhancement was applied to the composite. Although not a hard and fast rule, this was suggested for practical reasons, to avoid time-consuming re-work. Lastly, participant-witnesses were informed that they would make changes based on an initial composite, a face with features to match the verbal description prepared by the operator. This approach should result in a more effective composite as there would be a better context for selecting features, compared with selecting features using a face with 'default' features.

After answering any questions, operators used the verbal description from a participant-witness to assemble an initial composite. This face was presented to them to exchange, position and resize features as necessary. Witnesses were asked which feature should be worked on next, and operators carried out requests as required (as for all techniques). In this way, the order of feature selection was under the direction of the participant-witness. Operators carried out any changes required in the paint package: the E-FIT operator had available Picture Publisher; the FACES operator, Adobe Photoshop; and the PROfit operator, the paint package internal to PROfit as well as Adobe Photoshop.

Participant-witnesses producing a composite with the assistance of the sketch artist also began with feature selection. As with the other techniques, participant-witnesses were directed towards facial features that matched the verbal description. This time, materials were taken from the *FBI Facial Identification Handbook* (1988), the *Identikit Handbook Model II* (1960), and a selection of more recent hair styles (assembled by the artist). Participant-witnesses then focused on facial proportions to create a faint drawing with facial landmarks identified. Details were then 'fleshed out' using feature shapes adapted from the reference materials. As with the other techniques, the sketch artist worked under the direction of participant-witnesses to create the best possible likeness. The composite was sealed using a spray fixative.

Participant-witnesses constructing an EvoFIT first chose a hairstyle. The operator initially exported a reference image to PROfit, an average-looking face with short hair. Within PROfit, participants selected an appropriate hairstyle and the operator positioned and resized it as required. The image was then imported to EvoFIT, and the hairstyle extracted for use on all faces. Next, the EvoFIT operator presented a set of about 70 facial shapes (drawn randomly from an internal shape model) and participant-witnesses selected about six that best matched their target. They were then presented with a similar number of facial intensities or textures (drawn at random from the texture model) and again selected about six. Next, witnesses identified a face with the best overall likeness and the operator used the software to 'breed' together the selected shapes and textures. Participant-witnesses similarly selected from this 'evolved' set, with breeding applied. The process of selection and breeding continued until the witness identified the best possible likeness. A small utility was made available to allow features to be moved on demand (e.g. moving the eyes together) and Adobe Photoshop was used for any artistic enhancement (e.g. making the eyes lighter).

Debriefing involved as much detail of the project as required by the participant-witness.

Composite evaluation

Design

As mentioned earlier, composites produced using the modern computerized systems as well as artist's sketches appear good enough to be spontaneously named by third persons. Therefore, *composite naming* was our primary measure of composite quality. As research suggests that inspecting composites from different techniques interferes with naming (Frowd *et al.*, 2004), our participant-judges named composites from one system only. Thus, participant-judges saw a set of 10 PROfits, or a set of 10 E-FITs, etc. The resulting design was therefore between-subjects for construction technique.

Frowd *et al.* (2004) also asked their judges to name the target photographs after naming the composites. This has a two-fold benefit. Firstly, to check that participant-judges were sufficiently familiar with the targets. An *a priori* decision was applied which excluded those who did not name at least half of the targets (otherwise participants would not have been able to name many composites). Secondly, as participant-judges may not recognize all the targets, the composite naming level was calculated to reflect their familiarity with the target set: simply, this was the number of composites correctly named divided by the number of targets correctly named—a percentage correct score for each participant-judge (when multiplied by 100). The current work extends this notion by collecting target familiarity on a 5-point Likert scale, enabling an additional investigation of target familiarity on composite naming (i.e. one might expect participants to name more composites given greater familiarity with their targets).

Despite the forensic value of composite naming, composites tend not to be named very often—up to about 20% on average—and thus lie within a less sensitive region of the scale (i.e. performance is near floor level). Two supplementary tasks were therefore administered in the current research, *composite sorting* and *identification*. Both featured in previous research in this area. Composite sorting requires participant-judges to match composites to target photographs (in this case, 50 composites to 10 target photographs). To complete this task, participant-judges tend to compare individual facial features between the composites and the targets and therefore sorting represents a measure of feature quality (Frowd *et al.*, 2004). This appears to be a good proxy for naming (Davies *et al.*, 2000; Frowd *et al.*, 2004) and does not require a high degree of target familiarity (unlike naming). As participants inspect all composites, the design is within-subjects for composite technique.

The composite identification task involved judges comparing each composite with an array containing the relevant target photograph plus five similar-looking alternatives or 'foils'. A version with six items in the 'line-up' has been used previously to evaluate composites (e.g. Bruce *et al.*, 2002; Koehn & Fisher, 1997; Kovera *et al.*, 1997), although this method does not appear to have been formally compared against composite naming. We similarly adopted the six alternative forced-choice (6AFC) design, locating five similar-looking celebrity foils for each target (carried out by the first author based on a target's visual similarity). Note that all foils were different from each other (i.e. there were no repeats).

Participants

Composite naming was carried out by 80 participant-judges (16 participants $\times$ 5 conditions). There were 20 males and 60 females with an age range from 17 to 37 years ($M = 20.2$ years; $SD = 3.4$). All were psychology undergraduates at Stirling University who received a course credit for participation. A further group of 15 undergraduates received a course credit for the sorting task, 3 males and 12 females (from 18 to 35 years, $M = 22.6$ years, $SD = 5.9$). The identification task (line-up) comprised 90 volunteers (18 participants $\times$ 5 conditions) drawn from members of the Psychology department at Stirling University, and Open University students (attending Stirling University during Summer 2002). There were 62 females and 28 males (from 18 to 60 years, $M = 36.6$ years, $SD = 11.8$).

Procedure

Participant-judges for all three tasks were told initially that they would be evaluating composites of famous people constructed in a realistic study. Each person was tested independently. Those involved with naming were asked to provide a name for each composite where possible. If they were unable to recall a name, an unambiguous semantic description of the celebrity would be accepted—a procedure adopted elsewhere (e.g. Bruce, Hanna, Dench, Healey, & Burton, 1992). Participant-judges worked sequentially through a set of 10 composites (e.g. the 10 PROfits) in their own time and attempted to name them. After attempting to name the composites, participants were asked to rate each target photograph on our 5-point familiarity scale (refer to *Selection of target photographs*). Participants worked through the target photographs sequentially providing a familiarity rating score. The order of presentation for composites and targets was randomized for each person. Debriefing comprised as much information as required by the participant. This was the same for all tasks. Note that participants were discarded, and replacements found, if they did not know at least five of the target photographs (i.e. their familiarity rating was 2 or less on five or more targets); an *a priori* rule applied to ensure sufficient familiarity with the stimuli (see later for further details on target familiarity).

For the sorting task, the 10 target photographs were placed in front of each person. Participant-judges were given 70 composites[1] and asked to sort them into piles by placing each composite in front of a celebrity photograph (a forced-choice task). They were told that there were multiple composites of each target, though no mention was made of the number of repeats. Participant-judges worked sequentially through the pile of composites in their own time. The target and composite presentation order was randomized for each person.

Participant-judges involved in the identification task (line-ups) were presented with a booklet containing 10 composites, each with an associated target array (there were five booklets, one for each type of composite). Participants were told that each composite had been constructed from one of the photographs in the array, and they were required to select one. They were presented with each composite and, in their own time, selected a photograph from the corresponding array. Once again, composite presentation order was randomized for each person.

Results

Separate analyses were conducted for composite naming, sorting and identification. The following analyses compare performance for the five techniques (E-FIT, PROfit, sketch, FACES and EvoFIT).

Naming

For the five construction conditions, there were only 22 composites correctly named out of 800 possible naming responses (16 participants × 5 conditions × 10 composites)—only 2.8% correct. As detailed in Table 1, only 10 composites were correctly named by at least one person (out of the possible 50 constructed). Of these, five were sketches, two were PROfits and EvoFITs, and one was a FACES composite. In spite of this poor performance, a few of the composites were relatively well named. These included the sketch of Ian 'H' Watkins (31% naming), the FACES of Phillip Olivier (31%) and the EvoFIT of Kian Egan (19%).

Table 1. The number of composites correctly named by at least one person

Sketch	EvoFIT	PROfit	FACES	EFIT
5	2	2	1	0

Scores are out of 10, the number of composites constructed for each system.

After naming the composites, participant-judges provided a familiarity judgement for each target photograph (a score between 1 and 5). An analysis of the distribution of naming scores grouped by target familiarity revealed that seven composites were named for a target familiarity of 3, 13 composites for a familiarity of 4, and two composites for a

[1] *Our original design involved the construction of a second set of EvoFIT and FACES composites. These extra data were included in the sorting task but the results are not reported.*

familiarity of 5. We used a rating of 3 or more (3 = *Could say something about why he is famous, but not very confidently*) to indicate that a participant recognized a target, resulting in a mean target familiarity of 88%. These data were then used to calculate a composite naming measure based on familiar targets: by dividing the number of correctly named composites by the number of familiarity ratings in the category that had scored 3 or greater for each participant-judge. This resulted in a 'conditional' naming level of 3.3% overall (cf. 2.8% 'unconditional' naming level above). As can be seen in Table 2, sketches were named best (on average 8.1%), followed by EvoFIT (3.6%) and FACES (3.2%); E-FIT and PROfit were named by less than 2% of participant-judges.

Table 2. The three measures used to evaluate composites from the construction techniques investigated (sorted by naming scores)

Technique	Naming	Sorting	Identification
Sketch	8.1[a]	54.4[b]	46.7
EvoFIT	3.6	38.8	30.6
FACES	3.2	35.0	32.8
PROfit	1.3	40.6	40.6
E-FIT	0.0	42.5	60.0
Mean	3.2	42.0	42.1

Figures are percentage correct.
[a]Significantly greater than E-FIT and PROfit, $p < .05$.
[b]Significantly greater than all other techniques, $p < .05$.

The participant-judge naming scores (conditional) were subjected to a one-way independent-samples analysis of variance (ANOVA), which showed a significant effect of construction technique, $F(4, 75) = 4.98$, $p < .01$. *Post hoc* analysis (using Tukey HSD) indicated that sketch was greater than E-FIT and PROfit, $p < .05$. No other significant differences were found. As E-FIT and PROfit are very similar systems, two further *t*-tests were conducted with the E-FIT and PROfit naming data pooled (to increase statistical power). This revealed that both EvoFIT and FACES were better than E-FIT and PROfit, $t(76) > 2.40$, $p < .05$.

The mean participant target familiarity ratings ranged from a mean of 3.33 for those who saw the EvoFITs to 3.79 for those who saw E-FITs. These data were subjected to a second ANOVA, which was not significant for system, $F(4, 75) = 1.37$, $p > .05$, and suggests that naming between composite techniques was not affected by the participants' familiarity with the targets.

A point-biserial correlation between participants' (conditional) naming scores and mean familiarity was found to be non-significant and near-zero, $r(48) = .03$, $p > .05$. Unexpectedly, this suggests overall that familiarity with the target set did not influence the ability to name the resulting composites (though this may be due to very low overall naming).

Sorting

Overall, composites were sorted to an accuracy of 42% (refer to Table 2). Once again, sketch performed best (54.4%) followed by all other systems at about 40%. A sorting accuracy score (percentage correct) was derived for each participant and subjected to a one-way repeated-measures ANOVA. This was found to be significant for technique, $F(4, 60) = 4.22$, $p < .005$. Simple contrasts of the ANOVA confirmed that sketch was again significantly greater than all others, $p < .05$. Three *post hoc* (Bonferroni) tests were also run, resulting in equivalence for E-FIT and PROfit, $t(15) = 0.84$, $p > .0167$; E-FIT and FACES, $t(15) = 1.40$, $p > .0167$; and E-FIT and EvoFIT, $t(15) = 0.78$, $p > .0167$.

Identification (line-up)

For the identification task (line-up), target photographs were correctly identified 42.1% of the time (Table 2). E-FIT was best (60.0%), sketch second (46.7%) and PROfit third (40.6%). FACES and EvoFIT performed about 30%. Participant line-up scores (percentage correct) were subject to a one-way independent-samples ANOVA, and found to be significant for technique, $F(4, 85) = 25.37$; $p < .001$. *Post hoc* (Tukey HSD) testing revealed that E-FIT was significantly greater than all other techniques except sketch, $p < .05$, and sketch was greater than EvoFIT, $p < .005$. No other significant contrasts were found.

Other measures

To further evaluate the relationship between the dependent variables, point-biserial correlations were conducted. These (items) analyses revealed a medium level and significant correlation between (conditional) composite naming and sorting accuracy, $r(48) = .38$, $p < .05$, but a near zero correlation between naming and composite accuracy in the line-up task, $r(48) = .06$, $p > .05$. The correlation was also non-significant between the sorting and line-up tasks, $r(48) = .21$, $p > .05$.

A further correlational analysis was conducted between participant-witness's age and the quality of their composite, as measured by naming. This was found to be low and non-significant, $r(48) = .11$, $p > .05$, indicating that participant-witness age was not associated with composite quality. Similarly, a second correlation was conducted between the order of construction and composite naming. This was also found to be non-significant, $r(48) = .05$, $p > .05$, indicating that operators did not improve with practice: appropriate as all were previously experienced.

The average time taken to construct the composites varied considerably from about half an hour for FACES to over 2 hours for sketch. An ANOVA found that differences were significant across technique,[2] $F(4, 43) = 34.3$, $p < .001$. Tukey HSD tests revealed that EvoFITs ($M = 114$ minutes) and sketches ($M = 123$ minutes) took the same time to construct, $p > .05$, and were the slowest techniques, $p < .001$; E-FITs ($M = 67$ minutes) took the same time as PROfits ($M = 75$ minutes), $p > .05$, but were both slower than FACES ($M = 34$ minutes), $p < .005$.

Discussion

Five contemporary composite systems were evaluated under realistic conditions including a 2-day delay. The primary dependent measure, composite naming, suggested that composite quality was very poor in general: those produced from PROfit and E-FIT

[2] *A data collection error resulted in two missing data points for the PROfit system.*

(the two leading UK computerized systems), were correctly named only twice out of about 300 attempts. Composites from FACES, a relatively simple system, were better at 3.2%, and those from EvoFIT, an experimental system, were very similar (3.6%). The best named composites were produced by a sketch artist at 8.1%. These data did not appear to be influenced by participant-witness age nor operator learning, factors that might have weakened conclusions. The results also found a good match between composite naming and a simple sorting task, but a poor match between composite naming and a line-up task.

Very poor naming

Given the consistent finding that facial composites are named about 20% of the time—notably from E-FIT and PROfit systems (Brace *et al.*, 2000; Bruce *et al.*, 2002; Davies *et al.*, 2000; Frowd *et al.*, 2004)—why was naming so low? This result does not appear to be due to the target set employed, given that half the sketches were correctly named by at least one person (i.e. the targets were recognizable in one composite medium). Note that our procedure has been replicated while developing a 'parallel' interface to PROfit (Hancock & Frowd, 2003). This approach presents multiple features to a witness simultaneously, rather than one at a time (to encourage holistic face processing). A different experienced operator constructed a set of 20 composites using both types of interface, with a new set of similarly aged celebrity targets, but found very low naming levels (below 4%). Despite little difference between presentation formats, this work suggests that the results of the current study extend to another trained operator and a different set of target photographs inspected 2 days prior to construction.

It is conceivable that the poor quality composites were caused by limited verbal descriptions given by participant-witnesses. In the study, participants were required to describe each facial feature to allow an operator to locate appropriate features: the initial composite. Limited recall might hamper this process. However, feedback from operators suggests that descriptions were sufficient for an initial composite to be assembled and thus, although verbal descriptions may deteriorate over time—as found by Carson (2000) and Ellis *et al.* (1980) after 24 hours—limited recall did not appear to be an issue here. It is possible that verbal description production may hamper composite construction following shorter target exposures or longer periods of time, a notion that can be readily verified.

The very poor naming is perhaps best explained by a shift in cognitive processing. It is known that participants tend to remember a face as a set of features if a face perception task—such as composite construction—is anticipated (Olsson & Juslin, 1999). This 'feature-based' method of encoding is in contrast with a holistic encoding whereby observers perceive a face as a 'gestalt' or whole. It is known that feature-based encoding is best for feature-based construction (i.e. for E-FIT, PROfit and FACES; Wells & Hryciw, 1984), but that holistic encoding is best for face recognition and composites constructed via a sketch artist (Davies & Little, 1990; Wells & Hryciw, 1984). After a 2-day delay, it is likely that our participants' memory of a target is more of an impression (gestalt) than an analysis of features. Such a shift in cognitive processing would facilitate holistic over feature-based methods of construction, resulting in higher performance for sketch compared with PROfit, E-FIT and FACES (one would also expect composites from EvoFIT to be better named than PROfit, E-FIT and FACES, though this cannot be verified due to low values). In support of this notion, asking participants to describe a face (a feature-based task) also appears to encourage feature-based processing, as measured by a reduction in face recognition ability—known as a verbal overshadowing effect (VOE; e.g. Finger & Pezdek, 1999; Meissner & Brigham, 2001; Schooler & Engstler-Schooler, 1990). Interestingly, this deficit may be reversed following a short delay (24 minutes) between recall and recognition, apparently 'resetting' perceptual systems to their holistic state (Finger & Pezdek, 1999). It is perhaps the case that requesting witnesses to wait 2 days prior to recall—as in the current research—may similarly serve to encourage holistic face processing, yielding the observed benefits for sketched composites versus the other methods of construction (though witnesses would still describe a face, potentially interfering with recognition). We note that the current design does not allow an investigation of a VOE for composite construction, but acknowledge that it could be an interesting avenue of research.

EvoFIT and FACES

In addition to the main UK systems, the current work evaluated two other composite systems. Recall that EvoFIT uses a novel holistic approach, requiring witnesses to select whole faces that look like a suspect. EvoFIT faces have random characteristics initially, but through selection and breeding become more like a suspect. In earlier work (e.g. Frowd *et al.*, 2004), EvoFIT was worse than the UK alternatives, but with several issues resolved (mainly the software tools), performance was now comparable to E-FIT, PROfit and FACES. In fact, despite low naming levels, EvoFITs were named significantly better than the combined E-FITs and PROfits.

Despite the appeal of the approach, why were the EvoFITs not named better? One reason could be the use of a CI. As mentioned above, there is evidence that describing a face may interfere with subsequent recognition and, as EvoFIT is holistic (or recognition) based, verbal description production may interfere with constructing an EvoFIT (a similar argument applies to sketches). Based on work by Finger and Pezdek (1999) mentioned above, one approach would introduce a short delay between describing the face and constructing the composite. Alternatively, the CI could request more holistic information, a procedure found to facilitate recognition (e.g. Berman, Cutler, & Foos, 1991). Such a holistic CI would direct witnesses to consider the overall aspects of a face, perhaps asking; 'How would you describe the overall appearance? What kind of person do you think he is?' Indeed, there is evidence that making an inferential or trait judgment about a face promotes recognition and thus holistic processing (e.g. Devine & Malpass, 1985; Wells & Hryciw, 1984).

For the FACES system, given a relatively inexpensive product ($50), we anticipated that these composites would be worse than E-FITs and PROfits. Once again, due to the small values, strong conclusions based on composite naming are inappropriate—although, like EvoFIT, FACES composites were named significantly better than the combined E-FIT and PROfit composites. The sorting task did indicate that FACES were comparable to E-FIT, PROfit and EvoFIT.

There are two key differences between FACES, E-FIT and PROfit. Firstly, FACES contains a single mixed race and mixed gender 'library' of features, while the others have separate libraries. As such, it was anticipated that witnesses would be exposed to irrelevant features (e.g. presenting Asian features when white features were required), which would lead to witness confusion and poor composites. Secondly, FACES allows witnesses to select 'isolated' features—features out of the context of a whole face. Although this approach is contrary to previous research (e.g. Tanaka & Farah, 1993), it may allow witnesses to select facial features rapidly, and be a reason why shorter construction times were observed for FACES.

Proxies to naming

We now consider the two additional instruments included as possible alternatives or proxies to naming (valuable when naming levels are low). The sorting task required participants to assign composites to target photographs. It is a simple exercise requiring few participants (15 used here, cf. 80 for naming) but provided a good match with naming: sketches were sorted best (like naming) and, in spite of low naming levels, there was a medium-level correlation between the sorting and naming data. For the line-up task, six similar-looking photographs were presented for each composite. Unexpectedly, E-FITs performed as well as sketches, and there was a near-zero correlation between the line-up and naming data.[3] So, why might line-ups provide such a poor alternative to naming? One would suppose that as the arrays were selected to be similar to targets (as in past research, e.g. Bruce *et al.*, 2002), line-ups would be more of a holistic exercise than sorting, and therefore be more similar to naming (i.e. naming is a holistic task).

The mechanism underlying these data is not entirely clear. One problem is that, unlike naming, familiarity scores were not collected for the celebrity photographs and it is therefore difficult to assess differences in familiarity across tasks. In fact, differences in familiarity may introduce variability in the results. It is possible, for example, that participants in the line-up task were less familiar with the celebrities and therefore guided more by the ‘external’ facial features (e.g. Hancock, Bruce, & Burton, 2000), especially hair, producing less of a holistic task. It is possible, however, that sorting allowed participants to become more familiar with the stimuli set (sorting involved only 10 celebrity photographs; there were 60 for line-ups, as each array contained six celebrities) and allowed them to consider more facial features—especially the internal features of the face (e.g. eyes, nose, mouth) known to be important for naming—and thus served more as a holistic task.

In general, it would appear that, despite the simplicity of the task, composite sorting is a sensible proxy to naming; a result endorsed elsewhere (Davies *et al.*, 2000; Frowd *et al.*, 2004). At this stage, the use of line-ups seems more questionable, at least until further research has been conducted.

Generalization of results

The current work was careful to avoid allowing operators to construct more than one composite of the same target face, which resulted in a different operator for each composite system. This ensured that composites of the same identities did not influence each other—via the operator—and thus parallels police work where the independence of witness evidence must be maintained. Given research suggesting that operators themselves can affect the quality of a composite (Christie *et al.*, 1981; Davies *et al.*, 1983), might the current results be influenced by the skill of our operators? This is a general problem in facial composite research, one that we have attempted to address by selecting personnel who have undergone appropriate training, and who have constructed many

---

[3] *The low correlation between the naming and line-up data was not caused by the E-FITs, since this association is still poor with the E-FIT data excluded,* $r(38) = .17$, $p > .05$.

composites in the past. We note that even with one operator controlling more than one composite system, there is never complete certainly that operator skill is distributed equally. Arguably, the approach taken here is successful, given sensible naming and sorting data: E-FIT, PROfit and FACES are very similar systems and were found to perform equivalently despite being conducted by different operators (and despite recruiting different witnesses throughout, also known to exhibit large individual differences, e.g. Ellis *et al.*, 1978).

It is also important to consider that this study employed laboratory-based or 'mock' witnesses. Clearly the incentives for producing identifiable composites are very much higher in real life—the apprehension of a criminal compared with a small financial reward—and may lead to idiosyncratic behaviour. In support of this view, anecdotal reports from police operators to crime suggest that witnesses often repeatedly rehearse details about a suspect before attending an interview. Such a procedure may not only serve to promote a clearer mental image—as evidenced by a lift in face recognition ability following visual rehearsal (e.g. Sporer, 1988)—but also may encourage feature-based processing that may facilitate composite production (i.e. verbal recall and E-FIT/PROfit/FACES composite production are feature-based tasks). Thus, it would not be surprising to find better quality composites from E-FIT, PROfit and FACES when used more forensically. In general, our results appear to apply to witnesses who think little about the face of their 'assailant' prior to interview.

A second deviation from real life is the use of target photographs rather than a live target. Obviously, photographs provide a convenient medium to expose stimuli to participant-witnesses and may be used during composite evaluation (e.g. in naming and sorting tasks). However, might this choice of medium limit the generalization of results? It is clear that identifiable composites are possible with both live stimuli (i.e. a crime) as well as photographs (e.g. Brace *et al.*, 2000; Davies *et al.*, 2000; Frowd *et al.*, 2004). From the Shapiro and Penrod (1986) meta-analysis of 13 face perception studies, although static stimuli were found to produce less accurate identifications than live or videoed stimuli, the benefit was minimal (a shift of .09 to .14 in participant's criterion, B''). Given the general insensitiveness of the composite instrument, as evidenced by poor naming, it is unlikely that composite production would be affected by such a minor change.

Arguably, a third deviation from real life concerns our main dependent variable. Recall that participant-judges were asked where possible to produce a name for each composite. In police work, composites are accompanied by other information, including offender descriptions (e.g. race, age and height) and crime information (e.g. *modus operandi*). While it is tempting to assume that such information may be beneficial, there is evidence that person descriptions may be inaccurate (e.g. Cutshall & Yuille, 1989) and may therefore interfere with composite naming. The use of the un-cued naming method therefore provides an estimate of the identifiability of the composite alone.

Concluding remarks

As a final note, our results are likely to raise concerns with law enforcement agencies, in particular the very low naming rates, as these systems are typically employed several days after a crime—recall that a similar conclusion was reached by Koehn and Fisher (1997) using Mac-A-Mug Pro. In light of these findings, it would seem prudent for law enforcement agencies using computerized composite systems to interview witnesses sooner; although if this were not possible, the services of a sketch artist are likely to be more beneficial, in spite of the extra time required for construction. Our work also illustrates the need for further research, especially studies that attempt to improve composite quality, such as an exploration of mechanisms to facilitate feature-based processing for the E-FIT, PROfit and FACES systems.

Acknowledgements

This research was supported by a grant from the Engineering and Physical Sciences Research Council (GR/N09701). The authors would like to thank Claire Thomson-Bogner for her valuable assistance in composite data collection.

References

Association of Chief Police Officers (Scotland) Working Group (2000). *National working practices in facial imaging*. Unpublished manuscript.

Berman, G. L., Cutler, B. L., & Foos, P. W., (1991). Attempts to improve face recognition accuracy by inducing holistic processing and retrieval. In D. F. Ross & M. P. Toglia (Eds.), *Current trends in research on adult eyewitness memory and identification accuracy.* Symposium conducted at the meeting of the American Psychological Society, Washington, DC.

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

Bruce, V., Hanna, E., Dench, N., Healey, P., & Burton, M. (1992). The importance of "mass" in line-drawings of faces. *Applied Cognitive Psychology*, *6*, 619–628.

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.

Carson, D. (2000). *Cognitive interview: Research into the optimum time for conducting an interview.* Paper presented at the National ACPO working party for facial identification conference, Manchester, UK.

Christie, D., Davies, G. M., Shepherd, J. W., & Ellis, H. D. (1981). Evaluating a new computer-based system for face recall. *Law and Human Behaviour*, *2/3*, 209–218.

Cutler, B. L., Stocklein, C. J., & Penrod, S. D. (1988). An empirical examination of a computerized facial composite production system. *Forensic Reports*, *1*, 207–218.

Cutshall, J. L., & Yuille, J. C. (1989). Field studies of eyewitness memory of actual crime scenes. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 97–124). New York: Springer.

Davies, G., & Christie, D. (1982). Face recall: An examination of some factors limiting composite production accuracy. *Journal of Applied Psychology*, *67*, 103–109.

Davies, G., Milne, A., & Shepherd, J. (1983). Searching for operator skills in face composite reproduction. *Journal of Police Science and Administration*, *11*, 405–409.

Davies, G., van der Willik, P., & Morrison, L. J. (2000). Facial composite production: A comparison of mechanical and computer-driven systems. *Journal of Applied Psychology*, *85*, 119–124.

Davies, G. M., Ellis, H. G., & Shepherd, J. (1978). Face identification: The influence of delay upon accuracy of photofit construction. *Journal of Police Science and Administration*, *6*, 35–42.

Davies, G. M., & Little, M. (1990). Drawing on memory: Exploring the expertise of a police artist. *Medical Science and the Law*, *30*, 345–354.

Devine, P. G., & Malpass, R. S. (1985). Orienting strategies in differential face recognition. *Personality and Social Psychology Bulletin*, *11*, 33–40.

Ellis, H., Davies, G., & Shepherd, J. (1978). A critical examination of the photofit system for recalling faces. *Ergonomics*, *21*, 297–307.

Ellis, H. D. (1986). Face recall: A psychological perspective. *Human Learning*, *5*, 1–8.

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

FACES (2003). FACES version 3.0. Retrieved January 27, 2003, from http://www.iqbiometrix.com/products_faces.html

FBI facial identification handbook. (1988). Washington: FBI Graphic Design Unit—Special Projects Section—Laboratory Division.

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

Frowd C. D. (2002). *EvoFIT: A holistic, evolutionary facial imaging system*. Unpublished PhD thesis, University of Stirling, UK.

Frowd, C. D., Carson, D., Ness, H., Richardson, J., Morrison, L., McLanaghan, S., *et al.* (in press). A forensically valid comparison of facial composite systems. *Psychology, Crime and Law.*

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.

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.

Gibling, F., & Bennett, P. (1994). Artistic enhancement in the production of photofit likeness: An examination of its effectiveness in leading to suspect identification. *Psychology, Crime and Law*, *1*, 93–100.

Green, D. L., & Geiselman, R. E. (1989). Building composite facial images: Effects of feature saliency and delay of construction. *Journal of Applied Psychology*, *74*, 714–721.

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

Hancock, P. J. B. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments and Computers*, *32*, 327–333.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*, 330–337.

Hancock, P. J. B., & Frowd, C. D. (2002). Evolutionary generation of faces. In P. J. Bentley, & D. W. Corne (Eds.), *Creative evolutionary systems* (pp. 409–424). San Francisco: Morgan Kaufmann.

Hancock, P. J. B., & Frowd, C. D. (2003). *Parallel approaches to composite production: Recognition rather than recall*. Paper presented at the Society for Applied Research in Memory and Cognition, Aberdeen, UK.

Identikit Handbook Model II. (1960). Washington: Bangor Punta Operations.

Koehn, C. E., & Fisher, R. P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law*, *3*, 215–224.

Koehnken, G., Milne, R., Memon, A.,& Bull, R. (1994, March). *A meta-analysis of the effects of the cognitive interview.* Paper presented at the Biennial Conference of the American Psychology and Law Society, Sante Fe, NM.

Kovera, M. B., Penrod, S. D., Pappas, C., & Thill, D. L. (1997). Identification of computer generated facial composites. *Journal of Applied Psychology*, *82*, 235–246.

Krouse, F.L. (1981). Effects of pose, pose change, and delay on face recognition performance. *Journal of Applied Psychology*, *66*, 651–654.

Laughery, K., & Fowler, R. (1980). Sketch artist and identikit procedures for generating facial images. *Journal of Applied Psychology*, *65*, 307–316.

Laughery, K. R., Fessler, P. K., Lenorovitz, D. R., & Yoblick, D. A. (1974). Time delay and similarity effects in facial recognition. *Journal of Applied Psychology*, *59*, 490–496.

McNeil, J. E., Wray, J. L., Hibler, N. S., Foster, W. D., Rhyne, C. E., & Thibault, R. (1987). Hypnosis and Identi-kit: A study to determine the effect of using hypnosis in conjunction with the making of identikit composites. *Journal of Police Science and Administration*, *15*, 63–67.

McQuiston, D. E., & Malpass, R. S. (2000). *Use of facial composite systems in US law enforcement agencies*. Poster presented at the American Psychology-Law Society, New Orleans, LA.

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

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

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

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

Shepherd, J. W. (1983). Identification after long delays. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating witness evidence* (pp. 173–187). Chichester: Wiley.

Shepherd, F., Gibling, H.D., & Ellis, H.D. (1991). The effects of distinctiveness, presentation time and delay on face recognition. *European Journal of Cognitive Psychology*, *3*, 137–145.

Sporer, S. L. (1988). Long-term improvement of facial recognition through visual rehearsal. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (pp. 182–188). Chichester: Wiley.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 46A*, 225–245.

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

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, *15*, 525–536.

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology, 44A*, 671–703.

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

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

Wogalter, M., & Marwitz, D. (1991). Face composite construction: In view and from memory quality improvement with practice. *Ergonomics*, *22*, 333–343.

Received 14 February 2003; revised version received 23 July 2004