

A forensically valid comparison of facial composite systems
(Running Title: Evaluating Facial Composite Systems)

Charlie D. Frowd (1) – responsible for correspondence, etc. (cdf1@stir.ac.uk)

Derek Carson (2)

Hayley Ness (1)

Jan Richardson (3)

Lisa Morrison (4)

Sarah McLanaghan (1)

Peter Hancock (1)

(1) Department of Psychology
University of Stirling
FK9 4LA

(2) Department of Psychology
University of Abertay
DD1 1HG

(3) Independent Forensic Artist
Evidential Art
Market Harborough
Leicestershire
LE16 8YZ

(4) Department of Psychology
University of Leicester
LE1 7RH

Multiple Technique Composites in a “Forensically Friendly Format”

An evaluation of E-FIT, PROfit, Sketch, Photofit and EvoFIT composite construction techniques was carried out in a “forensically friendly format”: composites of unfamiliar targets were constructed from memory following a 3-4 hour delay using a Cognitive Interview and experienced operators. The main dependent variable was spontaneous naming and overall performance was low (10% average naming rate). E-FITs were named better than all techniques except PROfit, though E-FIT was superior to PROfit when the target was more distinctive. E-FIT, PROfit and Sketch were similar overall in a composite sorting task, but Sketch emerged best for more average-looking targets. Photofit performed poorly, as did EvoFIT, an experimental system. Overall, facial distinctiveness was found to be an important factor for composite naming.

Keywords: facial composite, memory, distinctiveness, witness

In a criminal investigation, there are many types of evidence. Evidence can be provided from the scene of the crime or from a witness (or victim). In the case of a witness, many demands are made. Through a Cognitive Interview, a witness is normally required to describe the events of a crime and the people concerned. A witness may also be involved in a line-up of similar-looking suspects, attempt to identify a suspect in a mugshot album or create a facial composite. The current research focuses on last of these, the construction of a composite, and explores the effectiveness of techniques available to law enforcement agencies.

A facial composite is a visual likeness of a perpetrator to a crime typically formed by the assembly of individual facial features. In a criminal investigation, a verbal description of the perpetrator is normally obtained from a witness. A technician skilled in the use of a composite system (referred hereafter as a composite *operator*) selects facial features that match this description. The witness then suggests changes to achieve an optimal likeness. A Sketch Artist follows a similar procedure, creating a composite by sketching using feature shapes selected by the witness.

In the past, the choice available to law enforcement agencies was limited to an artist composite or a manual composite system. Of the manual systems, there are two well-known types: Identikit and Photofit. The Identikit system, favoured in the US, uses facial features printed on acetate transparencies. The kit originally contained line drawings, but later, photographic elements (referred to as Identikit II). The Photofit system was adopted primarily in the UK and is similar to Identikit II. However, rather than acetates, facial features in Photofit are printed on jigsaw-like pieces that slot into a template. Whereas Photofit contains 855 features, Identikit II contains only 470 features, though transparencies can also be used in combination (Shepherd & Ellis, 1996). Both techniques provide a method of feature enhancement – such as adding of stubble or ageing wrinkles.

Early work by Ellis, Shepherd & Davies (1975) found that composites constructed by Photofit were only identified 12.5% of the time (from an array containing distractor faces). In Ellis, Davies & Shepherd's (1978a) participants constructed composites with different target modes (target-present or target-absent) and target exposures (15 seconds or 2.5 minutes); conditions under which differences would normally be expected in face perception tasks. These differences were not observed, which implied that there were deficiencies with the Photofit system. In the same year, Davies, Ellis & Shepherd (1978b) also found no significant difference between a composite created immediately and after 3 weeks. Another problem raised was that the feature demarcation lines present in Photofit appeared to interfere with identification (Ellis, Davies & Shepherd, 1978b). Concern was also expressed that Photofit did not contain an appropriate/sufficient selection of facial features (Davies, 1983a). Other research highlighted that selecting facial features in isolation to a whole face (the procedure used for Photofit) was undesirable and is at odds with the 'holistic' nature of face perception (e.g. Davies & Christie, 1982; Tanaka & Farah, 1993). Note that similar concerns have been voiced for the Identikit system (Green & Geiselman, 1989; Laughery & Fowler, 1980).

More recent computerised systems have attempted to overcome these problems (e.g. eliminating feature boundary lines, extending the range of features, presenting features for selection within the context of a whole face). One such system, Mac-A-Mug

Pro, runs on a Macintosh computer and contains “palettes” of sketch-like facial features that can be easily resized, moved and reoriented.

Cutler, Stocklein & Penrod (1988) found that composites constructed with this system were identified 49% of the time from an array containing many (60) distractors, suggesting good performance. However, composites were created with the target in-view and therefore cannot be seen as a valid test. Wogalter & Marwitz (1991) used an 8 second target exposure and found good matching ability (40%) for composites created from memory that, unlike Photofit, improved when composites were created with the target present. Unfortunately, they employed an unrealistically short target delay (20 minutes). Koehn & Fisher's (1997) participants constructed composites after 2 days. This time, only 4% of the targets were identified among few (5) distractors. A similar disappointing result was reported by Kovera, et al. (1997) with composites also constructed without continual reference to a target.

E-FIT is one of the leading composite systems in the UK. E-FIT is similar in design to Mac-A-Mug Pro, but contains photographed features to produce more realistic-looking composites. In their most ecologically valid method of construction, participants in Davies, van der Willik & Morrison (2000) constructed composites following a 1 minute target exposure using E-FIT and Photofit. A second group of participants attempted to name and sort the composites. No overall difference was found in either task, though the naming rate was low (17% for both E-FIT and Photofit). However, Davies et al. also report a good E-FIT naming rate (49%) for constructions made with the target present. In line with research mentioned earlier, Photofits were not named better when the target was present during construction. These results suggest that E-FIT has the potential to create good quality composites but this may be unattainable in the normal method of use (construction from memory).

Frowd, Hancock & Carson (2000) found a similar naming rate for E-FIT constructions from memory (18%), though participants creating the composites knew the identity of their target (i.e. a non-realistic method). Using PROfit, a similar UK composite system, Bruce et al. (2002) also found a comparable naming rate with composites constructed immediately after inspecting the target (19%).

In summary, it appears that Photofit and Identikit composites are poor quality and insensitive to the manipulations that normally give rise to change in the laboratory. The more flexible computerised systems such as Mac-A-Mug Pro and E-FIT are theoretically better, however performance remains disappointing when composites are constructed from memory. In this more realistic method of construction, the available research suggests that the modern systems fare no better than the older manual Photofit system. Note that there does not appear to be research published involving a significant target delay for E-FIT or PROfit. This may be an important factor given the poor quality Mac-A-Mug composites created after a more realistic delay (Koehn & Fisher, 1997; Kovera et al., 1997).

In Davies van der Willik & Morrison (2000), participants were recruited to construct a composite using both E-FIT and Photofit (a different photograph was inspected for each attempt). Constructions were first carried out following a 1 minute exposure to a target, and then the target was re-introduced after 20 minutes to allow an additional construction of a “target present” composite. In their work, 2 targets were presented per participant, one familiar and one unfamiliar. This resulted in 4 composites being constructed in a 1 hour session. The study was thus investigating the effect of

composite technique (E-FIT / Photofit), target presence (absent / present) and target familiarity (known / unknown). In composite naming and sorting tasks, they found that E-FIT and Photofit composites were equivalent, except when E-FITs were constructed of familiar faces in the presence of a photograph – the least forensically interesting condition. With witnesses working from memory, no significant difference emerged between E-FIT and Photofit – a curious result given that E-FIT was developed to overcome the deficiencies of Photofit.

Davies et al. draw attention to several limitations of their work. Firstly, only 20 minutes was allowed for composite construction from memory. This is a very short time and is unlikely to mirror any real life scenario. It is possible that participants may have been rushed, constructing 4 composites in an hour, and suggests that the systems were not used to their full potential. In a criminal investigation, it is normal practice for witnesses to make a single composite with the construction session being open-ended and terminated by the witness (not by the operator as in this study).

A second limitation mentioned by Davies and his colleagues is that artistic elaboration was not employed. There is evidence that enhancing the selected features to more closely match a target significantly improves the quality of a composite (Gibling & Bennett, 1985). For this reason, elaboration is used in police work. Without its use, it is difficult to generalize Davies and his colleagues' work beyond the laboratory. There are other problems limiting the scope of their results: they used neither a significant target delay (i.e. composites were constructed immediately after inspecting the target) nor a Cognitive Interview; both are present in a criminal investigation and may affect the quality of a composite (e.g. Frowd et al., 2000).

These limitations were addressed as part of our design that compared a number of composite techniques (including E-FIT and Photofit) as used in police work. We also believe that the general methodology adopted by many composite studies may lead to a problem of “operator contamination” (e.g. Davies et al., 2000; Ellis, Shepherd & Davies 1975; Ellis, Davies & Shepherd, 1978a; Koehn & Fisher, 1997), occurring when an operator constructs more than one composites of the same target. In this situation, an operator may be unknowingly influenced by composites constructed previously of the same target, thus producing similar-looking composites (and potentially reducing differences between experimental conditions). To avoid this undesirable effect, in UK criminal investigations where the same assailant is suspected, it is normal practice to recruit different operators (or artists). This is our approach and so different operators were employed for each technique.

Clearly, significant differences in the ability of operators are likely without proficiency in the necessary tools (e.g. basic software and artistic skills). Previous research has reported differences between novice and experienced operators with the Photofit kit (Davies, Milne & Shepherd, 1983). This suggests that only experienced personnel should be involved. It is also apparent that differences between operators may still emerge if substantial artistic elaboration is required, as would be the case for older-age targets. Care should thus be taken in the selection of targets to limit any enhancement for constructed composites.

The systems initially chosen for evaluation in this study were E-FIT and Photofit, to follow-up on Davies et al's work. Given the extensive forensic use in the UK (and Europe), and the apparent lack of a formal evaluation, PROfit was also included. Operators of these 3 systems were selected on the basis of their availability

(all our operators are authors of this paper). They were lab-based and had constructed composites for research projects in the past (using similar procedures to those adopted here). In terms of experience, the E-FIT and PROfit operators have attended an accredited facial composite course in the UK. The Photofit operator, while not having attended an accredited course, has constructed over 100 composites with ‘mock’ witnesses.

Given our contacts with the UK police, the study was also able to include a forensic police artist to construct a set of artist-composites. Our sketch artist is very experienced (in fact, the most experienced person in this study), completing a recognised course on facial composites in the late 1980s in the UK and subsequent composite training courses in America. She has been constructing composites in UK criminal investigations since initial training (late 1980s). Her input to the project has been invaluable, ensuring the design follows guidelines specified by the Association of Chief Police Officers (Scotland) (ACPO(S), 2000).

Involving these 4 techniques thus permits a comparison of past and present UK systems (though in the case of Photofit, McQuiston & Malpass (2000) have found occasional use in the US). We included a fifth system, called EvoFIT, under development at the University of Stirling, Scotland (Frowd, Hancock & Carson, 2000; Hancock, 2000; Hancock & Frowd, 2002). This novel computerised technique has yet to undergo a forensically valid test, and the current study appeared appropriate.

It was mentioned earlier that the current computerised systems attempt to capitalize on our ability to process faces as ‘wholes’ (as opposed to individual features) by presenting features within the context of a whole face. The new system takes this notion a stage further. EvoFIT faces are generated from a ‘holistic’ model using Principal Components Analysis, a statistical technique used to model changes in shape and intensity in a set of reference faces (currently about 70 young male faces). A face generated from this model has a set of holistic parameters, each with a global influence. For example, one parameter may widen a face, another may make it appear older.

Rather than selecting individual features - even in the context of a whole face - an EvoFIT is created by the selection and breeding of whole faces: witnesses are presented with a set of faces (about 70 in total) and choose those that most resemble their target (in fact, they first select facial shapes, then facial intensities or *textures*). EvoFIT then breeds together the preferred faces to produce another set. Witnesses next select from the ‘offspring’ faces, and these faces are bred together. With selection and breeding, the faces become more similar to each other and more similar to the target. When a witness identifies a face with a very good likeness, that image is saved as the ‘composite’. Therefore, a composite is created by ‘evolution’ (hence Evolution-FIT or *EvoFIT*). To assist in the evolution process, as witnesses sometimes suggest specific changes, software tools are provided which allow features to be manipulated on demand – e.g. making the eyes lighter or the face wider. For this new system, an EvoFIT operator was trained ‘in house’ and gained experience by constructing about 20 EvoFITs before the start of the study.

It is well documented in face perception that unusual, salient or distinctive faces are better remembered (Hancock, Burton & Bruce, 1996; Shapiro & Penrod, 1986; Valentine & Bruce, 1986; Valentine & Endo, 1992). One would expect that a composite of a more salient face would be superior to a composite of a more average face, a “distinctiveness” effect. Curiously, when tried with Identikit, composites of salient

targets were actually identified at chance level (Green & Geiselman, 1989), perhaps due to the limited feature repertoire of Identikit. Although Frowd et al. (2000) did not find a distinctiveness effect for E-FIT and EvoFIT constructions, composites of low distinctive targets were equivalent to composites of high distinctive targets. However, the target faces used were quite old ($M = 47$ years) and, as older faces are more distinctive, their entire target set was probably salient. Encouragingly, in a follow-up experiment, a distinctiveness effect was observed for EvoFIT with younger targets ($M = 27$ years), suggesting the potential of salient stimuli. We adopted this approach, using young targets with low and high distinctiveness.

It is predicted that composites from all techniques will exhibit a greater level of identification for distinctive-looking targets, although the effect size is likely to be diminished for Photofit due to a more limited feature repertoire. In general, it is expected that Photofit will be least able to express facial features, resulting in the worst composites.

For the two competing computerised systems in the UK, E-FIT and PROfit, we expect to find similar performance. The databases are very similar, sporting a large number of facial features and advanced software tools to manipulate and enhance facial features.

Frowd et al. (2000) also reported that EvoFIT, although able to create identifiable composites, did not perform as well as E-FIT. The authors argued that the target faces were generally too old for the EvoFIT database. Given a more age-appropriate target set, we anticipate equivalent performance for EvoFIT and E-FIT.

Although anecdotal reports may be found praising the prowess of a sketch artist (e.g. Taylor, 2001), the scarcity of relevant comparative lab-based research renders hypotheses difficult. Laughery & Fowler (1980) report that sketches were more identifiable than Identikit composites, though this may be due to the inflexibility of the line-based composite system employed. Ellis, Davies & Shepherd (1978a) found only a slight advantage for sketches over Photofit with constructions made from memory, but it is difficult to draw parallels from this work either as sketches were produced from witnesses themselves. In spite of work underscoring the variability of artist composites (Davies, 1986; Davies & Little, 1990; Laughery & Fowler, 1980), the sketch artist is the most flexible technique in our study with the potential to more closely reflect the demands of a witness. We therefore anticipate that sketches will perform as least as well as the other systems.

METHOD

The current study is a long overdue evaluation of UK composite systems (E-FIT, PROfit, Sketch, Photofit and EvoFIT). Constructions were carried out from a witness's memory of an unfamiliar face seen following a significant delay. A Cognitive Interview and artistic elaboration were used as part of the composite construction procedure. A different experienced operator was engaged in each technique and an operator worked on the same target only once (and remained blind to targets until composites had been constructed). Targets varied by facial salience and were chosen so that the composites would require minimal artwork (note that the same set of targets was used for each technique). Finally, in line with other research in this area, the targets were also chosen such that the resulting composites could be evaluated primarily by naming, arguably the acid test of a composite technique.

With different participants throughout, the construction of the composites was therefore a 5 x 2 between-subjects design for technique (E-FIT / PROfit / Sketch / Photofit / EvoFIT) and distinctiveness (high / low). To avoid potential learning effects, participants were recruited who had not constructed a composite in the past.

Selection of the target photographs

Famous faces were chosen as targets to enable our main dependent variable to be composite naming. As witnesses who construct composites do not know the identity of perpetrators, the same approach was followed, with our “witnesses” selecting an unknown famous face as their target. Our approach also used young (famous) adults as targets (individuals in their 20s and 30s) to limit ageing effects (crows feet, forehead wrinkles and under-eye bags), features that may exaggerate operator differences. Since most crimes are committed by young males (e.g. Goffredson & Polakowski, 1995), these stimuli appear appropriate.

Ten male celebrity targets were used, similar in number to other composite research (e.g. Cutler et al., 1988; Davies et al., 2000; Kovera et al., 1997). Based on ratings provided by an independent group of participants, half of the target photographs were selected as highly distinctive, the other half as more average. As distinctiveness estimates are also influenced by familiarity (Vokey & Read, 1992) - and witnesses in the current study do not recognize their target - we were only interested in distinctiveness ratings from faces that were not recognised. In practice, this involved collecting distinctiveness ratings from participants, then establishing which celebrities were known. This enabled distinctiveness scores to be analysed from celebrity faces reported as unfamiliar.

Static photographs were chosen as stimuli. In Shapiro & Penrod’s (1986) meta-analysis, an evaluation of 13 face perception studies, there was only a very minor change in participants’ behaviour (a small shift in criterion, B'' , from 0.09 to 0.14) between studies involving live (or videotaped) stimuli and those presenting static photographs. This suggests that a static stimulus is unlikely to affect composite quality.

Participants

Twenty-one volunteers were recruited from university staff, students and members of the public. There were 12 males and 9 females. Their age ranged from 23 to 70 with a mean age of 42.8 years ($SD = 13.7$).

Procedure

Twenty-two good quality colour photographs of young male celebrities were assembled (by the first author) using an extensive search on the Internet. All celebrities were in the range 20 to 40 years. This would allow participants to name the composites in the evaluation stage, when composites had been constructed. As far as possible, the targets were in a full-face pose and a neutral expression.

Participants were tested individually. They were told that they would be shown a number of photographs of celebrities, asked to imagine meeting each person at a railway station in amongst their peers (young, white males) and to rate these on distinctiveness from 1 to 7 (1 = *average, blend in to the crowd* and 7 = *very distinctive, stand out from the crowd*). Participants worked sequentially through the set providing an estimate of distinctiveness. Afterwards, the target photographs were re-presented and participants were asked to name them.

Further analysis concerned individual distinctiveness scores for celebrity faces not correctly identified. Five photographs with low average rated distinctiveness ($M = 3.1$, $SD = 1.5$) and 5 photographs with high average rated distinctiveness were extracted ($M = 5.2$, $SD = 1.5$); a significant difference using a t-test (an items analysis), $t(8) = 9.6$, $p < 0.001$. The mean age of the set was 29.4 years (refer to Table I for more details). One set for each operator (5 sets in total) were printed (in colour on the same high quality printer).

Table I

Target set used to construct the composites. The order is by increasing mean distinctiveness.

Celebrity	Occupation	Age (years)	Mean Distinctiveness	SD Distinctiveness
Michael Owen	Footballer	22	2.6	1.2
Damon Albarn	Singer	33	2.9	1.7
Stephen Gateley	Singer	24	3.3	1.4
Craig Phillips	Actor	30	3.6	1.4
Noah Wyle	Actor	30	3.7	1.5
Robbie Williams	Singer	26	4.3	1.8
Brad Pitt	Actor	38	4.9	1.6
Andre Aggasi	Tennis	31	5.0	1.6
David Beckham	Footballer	26	5.1	1.6
Noel Gallagher	Singer	34	5.8	1.2
Overall		29.4	4.1	2.4

Note. Distinctiveness is a rating on a 7-point scale (1 = *average, blend in to the crowd* and 7 = *very distinctive, stand out from the crowd*).

Constructing the composites

We opted for a longer, more realistic delay-to-interview than recent UK composite research (Bruce et al., 2002; Davies et al., 2000; Frowd et al., 2000). In the UK, police operatives follow guidelines from the Association of Chief Police Officers (Scotland) (ACPO(S)) and attempt construction within 24-36 hours, though in reality this can be much longer. A practical solution was chosen for operators and witnesses in this study with a 3-4 hour target delay, enabling targets to be presented in the morning and composites constructed in the afternoon; a tightly-controlled procedure that engaged witnesses for a day.

The 5 operators each recruited 10 witnesses to construct a set of composites (each operator used the same celebrities as targets). Witnesses identified a photograph of an unknown celebrity and inspected it for 1 minute. Three to four hours later, witnesses described the face using a Cognitive Interview and then worked with an operator to construct a composite using one of the 5 techniques.

Participants

Twenty males and 30 females were witnesses. Their age ranged from 18 to 81 years with a mean age of 40.1 years ($SD = 12.7$). They were drawn from university staff,

students and members of the public. Demographic information by technique is shown in Table II. Each person was paid £10.

Table II

Demographics of witnesses (participants constructing a composite)

	E-FIT	EvoFIT	Photofit	PROfit	Sketch	Overall
Males	4	7	3	4	2	20
Females	6	3	7	6	8	30
Mean age	40.7	39.4	31.6	37.4	55.6	40.9
SD age	7.5	9.8	4.3	11.7	17.3	13.3

Procedure

The basic procedure was the same for each construction technique. Participants who had not constructed a composite previously were identified. It was stressed that they would be acting as a “passive witness” and not placed in a stressful situation. It was also emphasised that a crucial aspect of the design was that “witnesses” did not know the identity of the target, to parallel real witnesses to crime.

Operators were responsible for presenting target photographs. They explained to participants that operators must not see any of the target photographs nor be told their names, as in real life. An envelope containing the targets was given to participants and then operators turned their back on them (to ensure targets were not seen by operators). Participants were instructed to remove a photograph from the envelope at random. If the celebrity in the photograph was recognised, they were asked to select another. If all photographs were known, the individual was thanked and dismissed (in practice, about three-times as many people were approached in order to recruit witnesses). Otherwise, the unfamiliar target photograph was inspected for 1 minute. Participants (now considered “witnesses”) were asked to report the target letter on the photograph (and operators recorded this code). Witnesses then placed the photograph in a second “used target” envelope.

Witnesses met with their operators 3 to 4 hours later. A procedure similar to a UK criminal investigation was followed for each technique, as outlined in ACPO(S) (2000). This was initiated by “rapport-building”, where operators chatted informally with witnesses. This was followed by an explanation of the process used to construct a composite: witnesses were told that a Cognitive Interview (CI) would first be carried out to recall details of their target’s face. Information recalled would then be used to construct a facial composite. Details of the CI were subsequently given: witnesses were told that shortly they would be asked to think back to when they saw the photograph and form a mental image of his face. When the image had been formed, they would then be asked to recall as many details about the face as possible. This is known as *free recall* and carried out in their own way and their own time, with minimal interruptions from operators. Witnesses were informed that operators would make notes during recall, useful for *cued recall*, the next phase. For this final part, information recalled about each feature is repeated to the witness (using the witnesses’ words). Witnesses would then be asked to focus on each feature in turn and try to recall more. Once any questions were answered, the CI was administered. The procedure described above (i.e. given to the witness) was followed, except that witnesses were asked to repeat the free recall stage after their first attempt (witnesses were not aware that a second recall was required).

The session moved on to the composite construction phase. For all techniques, the process used to construct a composite was detailed at the start and questions were answered as necessary. The process for constructing an E-FIT and PROfit composite was very similar. For these witnesses, it was explained that their composite system contained a large database of features for them to choose. They were shown how features could be selected, repositioned and resized. Witnesses were told that because facial features were cut from a limited set of photographs, only a general likeness may be possible. However, a paint package was available to improve the image, though it was normal for features to be selected before any artistic enhancement. Though not strict rule, this was suggested for practical reasons - to limit (time-consuming) re-work in the paint package when features were being exchanged.

After answering any questions, a composite was then constructed. Operators first used the witnesses' verbal description to prepare an 'initial' composite, consisting of features that matched the description. This image was presented to witnesses. They were encouraged to decide which feature to work on next and operators responded, exchanging, re-sizing and re-positioning features as necessary to achieve an optimal likeness. Witnesses were given the opportunity to enhance the likeness of their composite with a paint package, and operators carried out any necessary artwork. The PROfit operator used Adobe Photoshop version 5, the E-FIT operator used Picture Publishing version 5. Witnesses determined when the optimum likeness had been achieved, thus terminating the construction session. The composite was saved on disk.

A similar process was followed by the Photofit operator, initiated by the Cognitive Interview. Witnesses were introduced to the Photofit 'Visual Index', a set of reference photographs, used to select features. The Photofit operator assembled an initial composite, slotting features into a template. Once again, witnesses exchanged features in their order of preference to create the best possible likeness (though features could not be resized as for E-FIT and PROfit). When witnesses decided that the best possible likeness had been achieved, composites were recorded via a high quality camera, then scanned into a PC (for printing purposes).

Photofit feature boundary lines can interfere with recognition (Ellis et al., 1978b). To allow a fairer comparison with the other techniques, these lines were removed electronically in a paint package. This was achieved by concealing boundary lines with adjacent areas on the composite (using the copy, paste and resizing tools in Adobe Photoshop).

The sketch artist drew features shapes selected by a witness. Based on the verbal description, witnesses were directed towards the FBI Facial Identification Handbook (FBI, 1988) and the Identikit Handbook (Identikit, 1960). The former has photographs of faces showing features within the context of a face but do not reveal all the features on the whole face. The latter contains isolated features - some in sketch format, some photographic. Other reference materials were also available containing more recent hair styles (assembled by the artist).

Witnesses first selected appropriate feature shapes from these references. A sketch was constructed by first considering the facial shape and facial proportions (spacing between features). The outline of features was drawn lightly and modified as necessary from those selected. In general, the artist worked on groups of features, fleshing out the detail of the composite. When the best likeness had been achieved, the image was

preserved (using an aerosol fixative) and scanned into a computer (for printing purposes).

A two-fold process was necessary to construct an EvoFIT. As hair is not well-represented in EvoFIT, the PROfit system was used to locate an appropriate hairstyle. An average-looking face was first imported into PROfit and a hairstyle located. If artistic changes were necessary, the selected hairstyle was transferred into Adobe Photoshop for modification. The hairstyle was then transferred back into EvoFIT and faces were presented with this hair. Witnesses first selected about 6 facial shapes that looked like the target (from a set of about 70 example shapes), then a set of 6 facial colourings or *textures* (from about 70 example textures). The face with the best overall likeness to the target from the set was then selected, a so-called 'best-face'. These choices were next bred together to produce another set. Witnesses were given the opportunity to enhance their 'best-face', changing the size and/or position of features using a small utility; or altering image tone in Adobe Photoshop. The selection and breeding of faces continued until the witness achieved the best possible likeness, whereupon the face was saved to disk.

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

Construction Times

The time taken to conduct a Cognitive Interview and construct a composite was very similar for E-FIT ($M = 1$ hour 10 minutes, $SD = 15$ minutes) and PROfit ($M = 1$ hour 13 minutes, $SD = 14$ minutes). Sketch ($M = 2$ hours 15 minutes, $SD = 29$ minutes) and EvoFIT ($M = 2$ hours 35 minutes, $SD = 45$ minutes) sessions were also quite similar, though much longer. Photofit sessions were the shortest ($M = 47$ minutes, $SD = 11$ minutes).

COMPOSITE EVALUATION

As mentioned earlier, the primary dependent variable was composite naming. However, due to low naming rates traditionally found in this area, a simple sorting task was also administered. The design did not involve sorting with additional distracter items to parallel Davies et al. (and has been used elsewhere, e.g. Wogalter & Marwitz, 1991).

Naming

Participants who were naïve to the study attempted to name the composites. We acknowledge the "tip of tongue" phenomenon, a situation that finds a person temporarily unable to access a stored memory, and so allowed an unambiguous description; a procedure used elsewhere (Bruce, et al., 1992). For example, "Footballer, used to have a Mohican hairstyle, married to Posh Spice," was taken as a correct response for David Beckham. However, responses of "footballer" or "sportsman" was not acceptable.

Frowd and his colleagues report that mixing composites from several systems interferes with recognition, resulting in lower naming rates. To avoid this unwanted effect in the current work, each participant was presented with composites from one technique. Composite naming was therefore a 5 x 2 mixed design with technique a between-subjects factor and distinctiveness a within-subjects factor.

Participants

Twenty-six participants volunteered for each technique, a total of 130 (see Table III). Participants were university staff, students and members of the public. There were 58 males and 72 females. Their age ranged from 17 to 53 years with a mean age of 30.4 years ($SD = 10.0$).

Table III
Demographics of participants carrying out composite naming

	E-FIT	EvoFIT	Photofit	PROfit	Sketch	Overall
Males	8	8	12	19	11	58
Females	18	18	14	7	15	72
Mean age	33.3	31.2	33.2	29.5	24.8	30.4
SD age	8.9	7.5	6.7	9.1	7.0	8.3

Procedure

Participants were tested individually. They were presented with composites from one technique and told that each was created from the memory of a famous male and represented the likeness of a person rather than an identical image. They were asked to provide a name of the famous person depicted in the composite. Participants worked through each composite sequentially in their own time, providing a name where possible. No feedback was given regarding the accuracy of the response. The target photographs were then presented one-by-one for participants to name. The order of presentation for composites and target photographs was randomised after each person.

Results

Each target photograph was correctly named between 54% and 100% of the time. This large difference in target familiarity is a common problem with composite naming and suggests that the measure should be conditional on the number of targets known, an omission in recent work (e.g. Brace et al., 2000; Davies et al., 2000; Davies & Oldman, 1999). The “conditional” naming rate was calculated as the number of correctly named composites divided by the number of correctly named target photographs for each recogniser (expressed in percent). Table IV summarises these data by (high and low) distinctiveness. It can be seen that the highest rates were observed for E-FIT ($M = 19.0\%$) and PROfit ($M = 17.0\%$), then Sketch ($M = 9.2\%$), Photofit ($M = 6.2\%$) and EvoFIT ($M = 1.5\%$).

Table IV
Composite naming

	E-FIT	EvoFIT	Photofit	PROfit	Sketch	Overall
Low Distinctive	5.0	0.8	4.6	14.2	3.1	5.5
High Distinctive	33.1	2.3	7.7	19.8	15.4	15.7
Overall	19.0	1.5	6.2	17.0	9.2	10.6

Note. Scores are conditional naming rate, calculated by dividing the number of correctly named composites (in the low and high distinctive category) by the number of correctly named targets and averaged over participants (expressed in percent).

Table IV indicates an advantage for high distinctive composites. It turns out that a high distinctive composite was named best for each technique (an example can be seen in Figure I). There was only one well recognised low distinctive composite ($M = 46\%$), a PROfit of Michael Owen (the next best being a Sketch of Michael Owen, $M = 12.5\%$).

Figure I

Example of composites constructed of the Irish singer Noel Gallagher (a high distinctive target in this study).



Note. Each composite was constructed by a different witness. Shown are examples from E-FIT, EvoFIT, Photofit, Sketch and PROfit (left to right, top to bottom).

The conditional naming scores were subjected to a two-way analysis of variance (ANOVA). A significant effect was found for construction technique, $F(4, 125) = 12.0$, $p < 0.001$, and target distinctiveness, $F(1, 125) = 29.4$, $p < 0.001$. The interaction between these factors was also significant, $F(4, 125) = 6.8$, $p < 0.001$. Post hoc analysis (using Tukey HSD) for technique indicated that E-FIT was significantly greater than all other techniques, $p < 0.001$, except PROfit, and PROfit was greater than both EvoFIT, $p < 0.001$, and Photofit, $p < 0.01$. Simple-main effects for the interaction (LSD with $p < 0.05$) indicated a distinctiveness effect that was limited to E-FIT and Sketch. Whereas PROfit was best for low distinctive composites (caused by Michael Owen's PROfit, discussed later), E-FIT was superior for high distinctive targets in general. Also, for high distinctive targets, PROfit was better than both EvoFIT and Photofit, and Sketch was better than EvoFIT.

As one of the scores was an “outlier” - Michael Owen’s PROfit was named unexpectedly high for a low distinctive target - an items analysis is reported. The conditional naming rate *by-items* is the number of times each composite was correctly named divided by the number of times the corresponding target photograph was correctly named (equal to one score per composite, 50 in total). These scores were subjected to a two-way repeated-measures ANOVA (this time, both technique and distinctiveness were within-subjects factors). As before, a significant effect was found for construction technique, $F(4, 32) = 3.1, p < 0.05$, and target distinctiveness, $F(1, 8) = 7.9, p < 0.005$. This time, however, the interaction between these factors was not significant, $F(4, 32) = 1.3, p > 0.05$. These data suggest that distinctiveness is a consistent effect applying to all techniques, including PROfit.

Two further analyses were conducted to investigate potential sources of bias in the naming data. The first concerns witnesses, the second, operators. For the former, we examined whether the age of our adult witnesses, or their gender, would predict composite quality. Such an analysis was considered necessary, as differences in witness demographics may be significant. This can be seen in Table II: it turned out that Sketch witnesses were about 1 SD older than average; and there were also differences in the proportion of males-to-females (e.g. most Sketch witnesses were female, while most EvoFIT witnesses were male). A point-biserial correlation was found to be low and non-significant between composite naming and witness age, $r(48) = 0.11, p > 0.05$. To investigate possible witness gender effects on naming, the above ANOVA *by-subjects* for naming scores was repeated including gender as an additional factor. Collapsing over technique to increase statistical power, the ANOVA was significant for distinctiveness, $F(1, 46) = 8.3, p < 0.01$, but not gender, $F(1, 46) = 0.1, p > 0.05$ and there was no interaction, $F(1, 46) = 1.0, p > 0.05$. Therefore there is no evidence that neither witness age nor gender influenced composite naming.

The potential for operator improvement during the experiment was investigated. A point-biserial correlation was conducted between the conditional naming rate (*by-items*) and the order in which composites were constructed. This resulted in a low and non-significant correlation, $r(48) = 0.22, p > 0.05$, suggesting that there was no systematic bias in the order of construction. A lack of bias was expected as all operators were experienced (as detailed earlier) and hence were unlikely to have improved during the construction of composites in this study.

Sorting Task

This section investigates the quality of the composites using a simple sorting task. Participants were presented with all 50 composites and were required to sort them into piles, given the target photograph as reference. As witnesses were presented with composites from all techniques, the sorting task is therefore a 5 (technique) x 2 (distinctiveness) repeated-measures design.

Participants

Fourteen undergraduate students were recruited for the sorting task. There were 7 males and 7 females. Their age ranged from 17 to 35 years with a mean age of 22.1 years ($SD = 5.0$). Each person was paid £2.

Procedure

Participants were tested individually. They were given a pile containing all 50 composites and told that each had been constructed from one of 10 celebrities under realistic conditions. The 10 target photographs were laid out in front of them and told to match the composites to the target photographs. They were informed that there were multiple composites of each celebrity and therefore composites should be placed in a pile in front of each photograph (though details were not provided regarding the number of repeated composites for each target). Participants worked through the pile sequentially, in their own time, placing each composite in front of a celebrity. The order of the composites and the target photographs was randomised for each participant.

Results

Table V shows the number of composites correctly sorted by technique. Composites from E-FIT, PROfit and Sketch were sorted to an accuracy of approximately 70 to 80% compared with approximately 50% for EvoFIT and Photofit.

Table V

Composite sorting accuracy

	E-FIT	EvoFIT	Photofit	PROfit	Sketch	Overall
Low Distinctive	55.7	45.7	47.1	55.7	80.0	56.9
High Distinctive	92.9	54.3	50.0	88.6	81.4	73.4
Overall	74.3	50.0	48.6	72.1	80.7	65.1

Note. Scores are percent correct.

A two-way ANOVA for sorting accuracy (*by-subjects*) was significant for technique, $F(4, 52) = 21.0, p < 0.001$, and distinctiveness, $F(1, 52) = 19.8, p < 0.001$. The interaction was also significant, $F(4, 125) = 6.7, p < 0.001$. Simple-effects revealed a similar pattern to naming: E-FIT, PROfit and Sketch were equivalent overall, and better than EvoFIT and Photofit. For high distinctive composites, E-FIT was better than any other system except PROfit; PROfit was equivalent to Sketch; and all were better than EvoFIT and Photofit. For composites of low distinctive targets, sketches were sorted best. Lastly, for E-FIT and PROfit, composites of high distinctive targets were better than composites of low distinctive targets.

DISCUSSION

To summarize, this study was designed to compare 5 composite systems under realistic conditions. Results indicated that PROfit and E-FIT were overall equivalent on the naming task. E-FIT was named better than all other techniques (except PROfit). PROfit was the same as Sketch but greater than Photofit and EvoFIT. All techniques exhibited an elevated naming rate for composites of high distinctive targets (the items analysis indicated that this effect was consistent). Composite sorting revealed a similar pattern of results, though E-FITs were superior for high distinctive targets and sketches were superior for low distinctive targets.

The low mean naming rate of about 20% for E-FIT and PROfit fits into a larger body of research suggesting a similar level of performance (Brace, et al., 2000, Bruce et al., 2002, Davies, et al., 2000, Frowd, et al., 2000). Our supposition that E-FIT and PROfit would perform equivalently appears to be correct (even the construction times were almost the same). Both systems are very similar in design and use. There are differences: arguably the most salient being that the two databases contain different facial features; and the paint package used for artistic enhancement is internal to PROfit, but external to E-FIT. Overall, the data supports the notion that differences between the two systems are not sufficient to affect composite quality.

We were surprised to find a naming rate of only 9% for the artist's sketches, statistically less than E-FIT though the same as PROfit. However, if we pool the naming data for E-FIT and PROfit to increase statistical power, we find that Sketch is significantly less than these computerised systems, $t(76) = 3.1, p < 0.01$. In contrast, if we pool the E-FIT and PROfit data for the sorting task, we find an approaching significant advantage for Sketch, $t(13) = 1.71, p = 0.10$. This suggests that although sketched features are more accurate than E-FIT and PROfit, this benefit does not appear to be reflected in the identification rate. Note that this finding is unlikely to be caused by this person lacking artistic skills. Our artist is very experienced, having worked for about 15 years constructing composites in criminal investigations. Given her experience (she is by far the most experienced person with composites in the study), and the flexibility of the technique, we expected that sketches would be the most identifiable.

This result is curious: if the quality of the sketched features is superior to E-FIT and PROfit, but naming is worse, this suggests that the sketches are missing important information necessary for identification. An informal analysis (by the first author) noted that sketches created in the study tended to contain more detail for the face shape, hair, eyes, eyebrows and mouth; and less detail for the forehead, cheeks, chin and areas around and including the nose. With the possible exception of the nose, these areas of limited detail also tend to be omitted in descriptions of unfamiliar faces (Sporer, 1996).

Research has also found that representations containing less overall shading are not so well recognised. Firstly, there is a large body of research reporting worse identification when photographs are reduced to line drawings (e.g. Bruce et al., 1992; Davies, 1983b; Leder, 1996; Perrett et al., 1995). Secondly, poor identification has been reported from composites constructed using Mac-A-Mug Pro (Koehn & Fisher, 1997; Kovera, et al., 1997). The Mac-A-Mug system, although arguably not as flexible as a sketch artist, does nevertheless contain sketch-like facial features and could therefore exhibit similar problems.

These observations suggest that it might be valuable to include additional shading in a sketch for the major areas of the face covered by skin (e.g. forehead, cheeks, nose and chin). Unfortunately, as mentioned above, this information is unlikely to be conveyed by a verbal description. Our sketch artist has informally attempted to add extra artwork to make a sketch appear more lifelike. When this was carried out, witnesses were dissatisfied and the extra artwork had to be removed. It is clear that E-FIT and PROfit systems do not suffer from the problem of limited feature shading; these areas are present in the composite, and accompany the selection [made by the witness] of facial shape. These 2 systems also produced composites that were better named. Therefore, it might be prudent for sketch artists to direct witnesses to a broader range of

face shapes, similar to the PROfit and E-FIT databases, to facilitate a better impression of facial shading.

The current work attempted to limit operator differences using targets which did not require considerable artistic enhancement. The design appears appropriate, but did not take into account the fact that artist-composites themselves require artistic skill. This aspect was overcome by employing a suitably skilled police artist. The question is whether this study reflects the ability of one sketch artist or sketch artists in general? We believe the latter, as a combination of extensive experience and ongoing training (like our artist) is likely to result in optimal and consistent performance across artists. Clearly, to resolve this issue, it is necessary to compare several artists. Although Laughery & Fowler (1980) report such a comparison, further work might use artists with accredited training and long experience (their artists were not sufficiently trained or experienced).

The resulting Photofit composites appeared more in line with expectation. Despite the removal of feature boundary lines known to disrupt recognition (Ellis et al., 1978b), naming and sorting scores were relatively poor. Unfortunately, due to oversight, artistic elaboration was not carried out for Photofit and so inferences based on the data collected so far would be unfair. Ideally, it would be best to repeat the Photofit constructions with another operator and different witnesses (to avoid contamination and learning effects). However, due to discontinued use in the UK, another operator could not be identified. As an alternative to this, we compared the naming rate of Photofits with E-FITs following the removal of artwork. Such approach is sensible given that artistic elaboration tends to be carried out at the end of a composite session after the selection of facial features.

The E-FITs were printed again, this time without artistic elaboration. We recruited 18 new participants (undergraduates, naïve to the study) and presented each with a set of 10 composites for naming. This time we used 2 booklets, each containing 5 unelaborated E-FITs and 5 Photofits (order counterbalanced) to provide a more statistically powerful test. We found a conditional naming rate of 12.5% for E-FIT and 2.5% for Photofit. The difference in naming was significant, $t(17) = 2.9$, $p = 0.01$, providing clear evidence that E-FIT is better than Photofit.

Why then did Davies, van der Willik & Morrison (2000) fail to find a significant difference between E-FIT and Photofit with composites made from memory? We venture that their naming task, with multiple composites presented from the same target, could have potentially elevated naming rates and reduced the difference between experimental conditions. Another possibility is that the construction of 4 composites (with 2 different composite systems) in a 1 hour session was too demanding for a witness, reducing their performance and the constructed likeness. Which of these factors exerted the largest influence is unclear, and could be the focus of further work. It is apparent from the data reported here that using a methodology more closely aligned to a criminal investigation permits a reliable advantage to emerge for E-FIT.

The poor performance of Photofit fits within a significant body of research that has expressed concerns over its effectiveness (Christie & Ellis, 1981; Davies, Ellis & Shepherd, 1978; Ellis, Shepherd & Davies, 1975; Ellis, Davies & Shepherd, 1978a and 1978b). It would appear reasonable to infer that limitations in the features available remain important deficits of the technique (Davies, 1983). These limitations may be reflected in a shorter average construction time compared with E-FIT and PROfit (i.e. fewer features to search, less manipulation possible) - 47 minutes vs. 1 hour 10 minutes

(approx) - though this difference is perhaps due in part to a lack of artistic elaboration during Photofit sessions. It is curious to note that Photofit is still employed in the US (McQuiston & Malpass, 2000), a choice that would not appear to be optimal.

A poor naming rate was also found for EvoFIT (1.5%). This was considerably less than reported previously (10%, Frowd et al., 2000), in spite of a more appropriately aged target set here (in the previous work, the targets were believed to be too old for the EvoFIT database). In the current study, it was found that, despite training, the new operator experienced problems with the software. Although EvoFIT was designed to capitalize on face recognition ability, the system was rather complex to operate.

To create an EvoFIT, recall that a witness selects similar-looking faces that are bred together over a number of “generations”. Prior to starting each generation, a witness chooses the closest face from those selected - resulting in a set of “snap-shots” over time. Interestingly, pilot work showed a benefit of these “intermediate” images. We recruited 26 more participants (with similar demographics) and found that, when intermediate images were present along with the final EvoFIT, naming rose for the distinctive items to a level similar to PROfit and Sketch. It appears that this parallel format produces multiple triggers to recognition and may partly offset problems in operating the software (as earlier images may be superior). The parallel presentation of multiple images from one witness may be of benefit to composite construction in general, the same way as composites from multiple witnesses have proved valuable (Bennett, 2000; Bruce, et al., 2002). We are currently constructing a new set of EvoFITs with a more ergonomic version of the software.

In our work, we have tried to account for target and witnesses factors by using multiple targets and multiple witnesses (respectively) for each composite technique. Operator factors were limited by using young targets (to limit ageing effects) and “experts”: operators that were very experienced with their technique. We note that, with the exception of EvoFIT (as discussed above), the use of multiple operators worked well. The design has the advantage that operators do not construct more than one composite of the same target - thus avoiding operators being biased by previous composites - a matter of considerable forensic importance. However, with different operators in each condition, might they differentially influence composite quality? Past research does support this notion for Photofit operators (Christie et al., 1981; Davies, Milne & Shepherd, 1983). We assert that even with one operator, as in Davies et al. (2000), one can still never be sure that operator skill is equal across systems (i.e. an operator may be better using one of the techniques). Our design is thus a compromise - avoiding operator contamination at the expense of potentially elevating operator differences. The approach does appear at least in part to be valid, given ‘sensible’ results overall for E-FIT, PROfit and Photofit (i.e. E-FIT was the same as PROfit; E-FIT was better than Photofit). Arguably, to resolve this issue, one might seek to replicate the approach using a different experienced operator for a number of techniques, or the use of multiple experienced operators for one technique.

One issue (raised by one of the reviewers) concerns the notion of target familiarity. We supposed that a witness (honestly) claiming not to be familiar with a target was the same as the witness not having seen the target in the past. Clearly, as all our targets are famous, witnesses may have indeed seen them in the past, though not sufficiently often for recognition to have occurred during our study. Thus, it would appear more appropriate to say that our witnesses claimed not to be consciously familiar

with their target. Davies, van der Willik & Morrison (2000) investigated the effect of target familiarity and found no evidence that it affected Photofit or E-FIT constructions when a witness worked from memory (though familiarity did affect E-FITs when targets were available for continual inspection). We also note that Frowd et al. (2000) report E-FITs of known celebrities were named about the same as the current work. If familiarity were to be a factor, it is likely that naming rates in these other studies would have been substantially different (i.e. higher). Indeed, although familiarity may influence composite production, it is likely that the effect is too subtle to be measured with the composite instrument – a frequent complaint of the Identikit and Photofit systems (e.g. Ellis, Davies & Shepherd, 1978a; Ellis, Shepherd & Davies, 1975; Laughery & Fowler, 1980). Clearly, this issue could be resolved in follow-up work using non-celebrity faces, though collecting naming data may be more difficult (especially when a large number of composites may be involved).

In general, research appears to be consistent and converging on a naming rate of about 20% for E-FIT and PROfit (Brace et al., 2000; Bruce et al., 2002; Davies et al., 2000; Frowd et al., 2000). In our work, artist composites were named about half as often. Does this mean that, in a criminal investigation, suspects will be identified from composites with these naming rates? In real life, composites are normally accompanied by context information including a description of the perpetrator, information about the crime scene and any other salient information (e.g. method of operation and patterns of offending). This information may serve to elevate the naming rate, but also the number of false alarms, though research in this area is limited and tends to focus on comparing composites with verbal descriptions (Christie & Ellis, 1981) rather than potential additive effects. In general, our research reflects identification from composites given minimal contextual information (i.e. famous faces). A role of future research then might explore methods through which the police distribute composites (both internally and to the public) to determine the effect of context on composite naming.

Another valuable finding was that for the 5 techniques investigated, composites of high distinctive targets were named significantly better than composites of low distinctive targets. This effect was not consistent for composite sorting, emerging for just E-FIT and PROfit. Of course, sorting is arguably less important than naming, revealing more about the quality of individual features (participants sort largely by comparing features between targets and composites). In particular, our results suggest that the quality of sketched features is independent of target salience, unlike other current UK systems (though it could also be argued that the sparser representation of sketches renders them easier to sort – i.e. less information for comparison, thus reducing the cognitive load). Our results also suggest that sorting is a good proxy to naming, especially for the current UK computerised systems (as found by Davies, van der Willik & Morrison (2000)).

As far as we are aware, our work is the first to demonstrate a distinctiveness effect with facial composites. Recall that Green & Geiselman (1989) reported a *typicality* effect with Identikit, with constructions better from more average-looking targets. It is likely then that the modern computerised systems, with more features and flexibility, can take advantage of salient stimuli. In general, an advantage for constructing salient stimuli is not surprising. It is an important concept in face perception backed by considerable research (e.g. Bruce, 1988; Hancock, Burton & Bruce, 1996; Shapiro & Penrod, 1986). Recall that Frowd et al. (2000) did not report a

distinctiveness effect for E-FIT with composites constructed from memory. The authors supposed that targets used were inherently distinctive due to relatively high target age (about 50 years). This hypothesis appears to be correct, since our lower aged target set did reveal a distinctiveness effect. This also suggests that facial distinctiveness may be less important for composites constructed of older assailants.

REFERENCES

ACPO(S) (2000). National Working Practices in Facial Imaging. Association of Chief Police Officers (Scotland) Working Group.

Bennett, P. (2000). The use of multiple composites in suspect identification. *Proceedings of the 3rd UK national conference on cranio-facial identification*. Manchester, May 2000.

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*. Krakow: Institute of Forensic Research Publishers.

Bruce, V. (1988). *Recognising faces*. LEA.

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*.

Bruce, V. & Young, A. (1998). *In the eye of the beholder: the science of face perception*. Oxford University Press.

Caldwell, C., & Johnston, V.S. (1991). Tracking a Criminal Suspect Through "Face-Space" with a Genetic Algorithm, *Proceedings of the Fourth International Conference on Genetic Algorithms*, 416-421. Morgan Kaufmann Publishers.

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, 209-218.

Christie, D., & Ellis, H. (1981). Photofit constructions versus verbal descriptions, *Journal of Applied Psychology*, 66, 358-363.

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.

Davies, G.M. (1983a). Forensic face recall: the role of visual and verbal information. In S.M.A. Lloyd-Bostock & B.R. Clifford (Eds.). *Evaluating witness evidence*. London: John Wiley & Sons Ltd., Ch. 6.

Davies, G.M. (1983b). The recognition of persons from drawings and photographs. *Human Learning*, 2, 237-249.

Davies, G.M. (1986). Capturing likeness in eyewitness composites: the police artist and his rivals, *Med. Sci. Law*, 26, 4, 283-290.

Davies, G.M. (1996). Children's identification evidence. In S.L. Sporer, R.S. Malpass & G. Koehnken (Eds.). *Psychological issues in eyewitness identification*. Hillsdale, NJ: Lawrence Erlbaum. Ch. 10.

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

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, 1, 35-42.

Davies, G.M., & Little, M. (1990). Drawing on memory: exploring the expertise of a police artist, *Med. Sci. Law*, 30, 4, 345-354.

Davies, G. M., Milne, A., & Shepherd, J. (1983). Searching for operator skills in face composite reproduction. *Journal of Police Science and Administration*, 11(4), 405-9.

Davies, G.M., & Oldman, H. (1999). The impact of character attribution on composite production: A real world effect? *Current Psychology: Developmental, Learning, Personality, Social*. 18 (1), 128-139.

Davies, G.M., van der Willik, P., & Morrison, L.J. (2000). Facial Composite Production: A Comparison of Mechanical and Computer-Driven Systems, *Journal of Applied Psychology*, 85, 1, 119-124.

Ellis, H., Davies, G.M., & Shepherd, J. (1978a). A critical examination of the photofit system for recalling faces, *Ergonomics*, 21, 4, 297-307.

Ellis, H., Davies, G.M., & Shepherd, J. (1978b). Remembering pictures of real and 'unreal' faces: some practical and theoretical considerations, *British Journal of Psychology*, 69, 467-474.

Ellis, H. Shepherd, J., & Davies, G.M. (1975). Use of photo-fit for recalling faces. *British Journal of Psychology*, 66, 29-37.

FBI (1988). *FBI Facial Identification Handbook*. FBI Graphic Design Unit - Special Projects Section - Laboratory Division.

- Finger, K. & Pezdek, K. (1999). The effect of the cognitive interview on face identification accuracy: release from verbal overshadowing. *Journal of Applied Psychology*, 84, 3, 340-348.
- Frowd, C.D. (2002). EvoFIT: A Holistic, Evolutionary Facial Imaging System. *Unpublished PhD Thesis*, University of Stirling.
- Frowd, C.D., Hancock, P.J.B., & Carson, D. (2000). EvoFIT: A Holistic, Evolutionary Face Identification Technique. Presentation to the *Association of Chief Police Officers (Scotland)*. Manchester.
- Gibbling, 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.
- Goffredson, M.R., & Polakowski, M. (1995). Information retrieval: reconstructing faces. In N. Brewer & C. Wilson (Eds.). *Psychology and Policing*. Hillsdale, NJ: Lawrence Erlbaum. 101-117.
- 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.
- Hall, D.F. (1976). Obtaining eyewitness identification in criminal investigations: applications of social and experimental psychology (Doctoral dissertation, Ohio State University, Columbus). *Dissertation Abstracts International*, 37, 2569B.
- Hancock, P.J.B. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments and Computers*, 32-2, 327-333.
- Hancock, P.J.B., Bruce, V. and Burton, A.M. (2000). Recognition of unfamiliar faces. *TICS*, 4-9, 330-337
- Hancock, P.J.B., Burton, A.M., & Bruce, V. (1996). Face processing: human perception and principal components analysis. *Memory and Cognition*, 24, 26-40.
- Hancock, P.J.B. and Frowd, C.D. (2002). Evolutionary generation of faces, in Bentley, P.J. and Corne, D.W. (Eds.). *Creative evolutionary systems*. Morgan Kaufmann. Ch. 16.
- Identikit (1960). *Identikit Handbook Model II*. Published 1960:1975:1983. Bangor Punta Operations Inc.
- Koehn, C.E., & Fisher R.P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law*, 3, 215-224.
- Kovera, M.B., Penrod, S.D., Pappas, C., & Thill, D.L. (1997). Identification of computer generated facial composites. *Journal of Applied Psychology*, 82(2), 235-246.

Laughery, K.R., Duval, C., & Wogalter, M.S. (1986). Dynamics of facial recall, in Ellis, H.D., Jeeves, M.A., Newcombe, F., & Young, A. (Eds.). *Aspects of face processing*. Dordrecht: Martinus Nijhoff. 373-387.

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

Leder, H. (1996). Line drawings of faces reduce configural processing. *Perception*, 25, 355-366.

Mauldin, M., & Laughery, K. (1981). Composite production effects upon subsequent facial recognition, *Journal of Applied Psychology*, 66, 351-357.

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

Perrett, D., Benson, P.J., Hietanen, J.K., Oram, M.W., & Dittrich, W.H. (1995). When is a face not a face? In R. Gregory, J. Harris, P. Heard & D. Rose (Eds.). *The Artful Eye*, Oxford University Press.

Rakover, S.S., & Cahlon, B. (1996). To catch a thief with a recognition test: the model and some empirical results. *Cognitive Psychology*, 21, 423-468.

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

Shepherd, J., & Ellis, H. (1996). Face recall - methods and problems. In S.L. Sporer, R.S. Malpass & G. Koehnken (Eds.). *Psychological issues in eyewitness identification*. Hillsdale, NJ: Lawrence Erlbaum. Ch. 5.

Shepherd, J.W., Ellis, H.D., & Davies, G.M. (1982). *Identification evidence: a psychological evaluation*. Aberdeen: Aberdeen University Press.

Sporer, S.L. (1996). Psychological aspects of person descriptions. In S.L. Sporer, R.S. Malpass & G. Koehnken (Eds.). *Psychological issues in eyewitness identification*. Hillsdale, NJ: Lawrence Erlbaum. Ch. 4.

Tanaka, J.W., & Farah, M.J. (1993). Parts and wholes in face recognition, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 46A, 2, 225-245.

Taylor, K.T. (2001). *Forensic art and illustration*. CRC Press LLC.

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

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(3), 291-302.

Wells, G.L., & Hryciw, B. (1984). *Memory & Cognition*, 12, 4, 338-344.

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

Yu, C. & Geiselman, R.E. (1993). Effects of constructing Identi-kit composites on photospread identification performance. *Criminal Justice and Behavior*, 20, 280-292.