An evaluation of US systems for facial composite production

C.D. FROWD*†, D. MCQUISTON-SURRETT‡, S. ANANDACIVA#, C.G. IRELAND# and

P.J.B. HANCOCK#

†University of Central Lancashire, UK

‡Arizona State University, US

#University of Stirling, UK

*Corresponding author: Charlie Frowd, Department of Psychology, University of Central Lancashire, Preston PR1 2HE, UK. Email: cfrowd@uclan.ac.uk. Phone: (01772) 893439.

Abstract

Witness and victims of serious crime are normally requested to construct a facial composite of a suspect's face. While modern systems for constructing composites have been evaluated extensively in the UK, this is not the case in the US. In the current work, two popular computerized systems in the US, FACES and Identikit 2000, were evaluated against a 'reference' system, PRO-fit, where performance is established. In Experiment 1, witnesses constructed a composite with both PRO-fit and FACES using a realistic procedure. The resulting composites were very poorly named, but the PRO-fit emerged best in 'cued' naming and two supplementary measures: composite sorting and likeness ratings. In Experiment 2, PRO-fit was compared with Identikit 2000, a sketch-like feature system. Spontaneous naming was again very poor, but both cued naming and sorting suggested that the systems were similar. The results support previous findings that modern systems do not produce identifiable composites.

Keywords: facial composite; witness; evaluation; interview; crime

Acknowledgements

This research was supported by a grant from the Engineering and Physical Sciences Research Council (GR/N09701). The authors would like to thank Imran Kirkland, Alex McIntyre, and Gina Erickson for their valuable assistance in data collection.

1. Introduction

Facial composites are visual likenesses of human faces. They are normally constructed by witnesses and victims, who first describe the facial appearance of a suspect, and then select individual facial features from a kit of parts: hair, face shape, eyes, nose, mouth, etc. The earliest technique for constructing composites involved a sketch artist, a person skilled in portraiture, drawing the face by hand using pencils or crayons. Techniques were devised in the 1960s for use by those less artistic, and included Photofit, which was popular in the UK, and Identikit, in the US (e.g. Ellis et al. 1975, 1978, Laughery and Fowler 1980). Considerable research has been conducted on their evaluation (see Davies and Valentine 2006, for a review), the results of which have led to computerized systems which the police now use. Modern examples include E-FIT and PRO-fit in the UK, and FACES, Identikit 2000, CompuSketch, Mac-a-Mug, and SuspectID, in the US.

The UK systems have been subjected to a number of formal evaluations. These studies have found that E-FIT and PRO-fit produce composites that are correctly named about 20% of the time when participant-witnesses attempt construction either immediately or a few hours after seeing a target face (Brace et al. 2000, Bruce et al. 2002, Davies et al. 2000, Frowd et al. 2004, 2005b, 2007a, 2007b). Unfortunately, when participant-witnesses are required to wait two days prior to construction, a situation typical of real witnesses, composite naming normally falls to a few percent correct at best (e.g. Frowd et al. 2005a, 2005c, 2007b).

In spite of a greater range of techniques available in the US, evaluations thereof are rare; the authors are only aware of one: Frowd et al. (2005a). In this work, UK and US techniques were compared under a realistic two day delay. Along with the three UK techniques in current police use – E-FIT, PRO-fit and sketch artist – a system called EvoFIT was evaluated. EvoFIT is a new computerized technique that works by the selection and breeding of whole faces (e.g. Frowd et al. 2004, 2006a, 2006b, 2007b). The fifth system was FACES 3.0. This is a popular, but inexpensive, computerized method from the US (originally priced at $50, compared with thousands of dollars for PRO-fit and E-FIT). The study found that the composites from a sketch artist were correctly named 8% of the time, but the other systems were worse ($M < 4\%$). Analyses of these data were hampered by low values, but a supplementary measure was employed, referred to as composite sorting, which also suggested that the computerized systems were equivalent.

The current work sought to investigate the two most popular US computerized systems (McQuiston-Surrett et al. 2006) using a more powerful design than Frowd et al. (2005a). To do this, witnesses here each constructed two composites, one from a US system and one from PRO-fit, a ‘reference’ (UK) system that has been evaluated extensively (Frowd et al. 2005a, 2005b, 2005c, 2006b, 2007a, 2007b, 2007c). In Experiment 1, PRO-fit and FACES were compared; in Experiment 2, a similar comparison was made between PRO-fit and Identikit 2000.

2. Experiment 1 – PRO-fit versus FACES

Experiment 1 compared PRO-fit and FACES. A more powerful experimental design was used than in Frowd et al. (2005a). Firstly, laboratory-witnesses constructed one composite with each of these systems in order to allow the system itself to be a within-subjects factor and reduce experimental variability. Secondly, the design employed only one 'operator' to control the composite software. It is known that operators can exert a significant influence on composite quality (Christie et al. 1981, Davies et al. 1983), and the use of a single operator may further help to limit variability in system use.

Both FACES and PRO-fit contain a large number of facial features for a witness to select. There are however several notable differences. While PRO-fit contains features from different races in separate databases, such features are combined in FACES. Consequently, witnesses using FACES are sometimes inappropriately presented with features from another race; for example, Chinese noses may be shown when White Caucasian examples are required. Secondly, feature selection with FACES is carried out in isolation from a whole face, a procedure found not to be optimal (e.g. Davies and Christie 1982, Tanaka and Farah 1993); features in most other systems (e.g. PRO-fit, E-FIT) are switched in and out of the presented face. Thirdly, while features may be resized and positioned freely with PRO-fit, these functions are more limited with FACES. Lastly, there is an artwork package available within PRO-fit for enhancing the appearance of facial features; but no such program is available for FACES. Given the importance of being able to artistically enhance a composite (e.g. Gibling and Bennett 1994), the current study employed Adobe Photoshop for use with FACES. As the first three of these differences favour the PRO-fit system, it was predicted that FACES composites would be of worse quality than those produced by PRO-fit.

Two stages were required to carry out such an evaluation. In the first stage, participant-witnesses inspected a target face and then constructed a composite using both systems, one at a time. In the second stage, the quality of the composites was evaluated by third persons. The design employed famous faces as targets so that the resulting composites could be primarily assessed by spontaneous naming, arguably the most forensically interesting evaluation measure, although supplementary tasks were also administered.

2.1. Composite construction

2.1.1. Method. The general procedure of Frowd et al. (2005a) was repeated to construct each composite, a design which matches police procedures used with witnesses in both the UK and US (although there is some evidence that procedures in the US may be more variable than in the UK, presumably due to more variability in system training; see McQuiston-Surrett et al. 2006). Thus, after inspecting an unfamiliar celebrity target, participant-witnesses waited two days and then underwent a Cognitive Interview (CI), a procedure known to assist witnesses' recall (e.g. Geiselman et al. 1986), and constructed a composite using both FACES and PRO-fit. To do this, they interacted throughout with a composite operator. The CI used included three stages: rapport building, free recall and cued recall, described in more detail below. The verbal description was then used by the

operator as part of composite construction, first with one system, then with the other (order randomised).

As mentioned above, the skill of the operator can affect composite quality. To obtain an operator with roughly equal experience in PRO-fit and FACES systems, an experimenter with little experience of facial composites systems received training ‘in house’ and then practiced extensively and equally with both systems. He was also given training in the application of the CI.

2.1.2. Participants. Ten staff and students at Stirling University were each paid £10 ($15) to be witnesses. There were four males and six females with an age range from 19 to 52 years (M = 32.2 years, SD = 10.3).

2.1.3. Materials. The targets were photographs of 10 young celebrity male faces used by Frowd et al. (2005a). These faces were depicted without glasses and, as far as possible, in a front face pose with a neutral expression and minimal facial hair. They consisted of actors (Ben Affleck, Matt Damon, Jeremy Edwards, Joshua Jackson, Philip Oliver, and James Redmond) and pop singers (Kian Egan, Mark Feehily, Ronan Keating, and Ian 'H' Watkins). The mean age of the set was 26.3 years and they were well known by our undergraduates, who would later name the composites.

The experiment used PRO-fit version 3.1, marketed by ABM-UK (http://www.abm-uk.com/uk/index.asp); FACES 3.0, marketed by IQ Biometrix (http://www.iqbiometrix.com); and Adobe Photoshop 5.0.

2.1.4. Procedure. Participants were tested individually and made two visits to the laboratory. In the first visit, they studied a target photograph of a male celebrity, who was unrecognised by them. Participants were given an envelope containing the target photographs and asked to select one at random. If the famous face was reported familiar, the photograph was returned and another selected. If all photographs were recognised, participants were thanked and dismissed (this occurred twice). Otherwise, they were given one minute to look at the photograph with the knowledge that a composite would be required in two days time. All target photographs were used once. This procedure was carried out without the operator seeing the celebrity targets, so that he was unable to give inadvertent assistance during composite construction.

Participant-witnesses returned to the laboratory two days later and were first given a Cognitive Interview. This was initiated by a rapport building stage whereby the operator and witness chatted informally for several minutes. An overview of the session was then given to explain that procedures used would follow those of real witnesses, and so they would first describe the target’s face using a Cognitive Interview, and then construct a composite. Witnesses were invited to ask questions when necessary.

The operator provided an overview of the Cognitive Interview. It was explained that they would first be required to recall as much detail as possible of the target’s face. This was called free recall and would be carried out with minimal interruption from the operator. Witnesses were encouraged to describe in their own time and in the order of their choice. A cued recall phase would follow, with the given description repeated for each facial feature

and a prompt made for further recall. When ready, the Cognitive Interview was administered using this procedure.

The session moved on to composite construction. The order of system use (PRO-fit or FACES) was randomised such that each was used first half of the time. Witnesses were introduced to the system to be used first; no mention was made that more than one composite would be attempted. It was explained that the first stage was for the operator to enter the verbal description into the composite system. This would allow an initial composite to be assembled, a face with features to match the description. Participant-witnesses would then be able to exchange, size and position features in this face to obtain the best possible likeness. A paint package was available to improve the likeness of any feature, although this would normally be deployed towards the end of the session. A short demonstration was given regarding the selection and manipulation of features.

When ready, a composite was constructed as described. Witnesses worked at their own pace, were given the opportunity to work on features of their choice – although this was normally hair and face shape initially – and decided when the best likeness had been achieved. The procedure was the same irrespective of the system used, with two exceptions. Firstly, FACES presented both a complete composite face as well as a set of isolated features from which witnesses could select; PRO-fit witnesses were only presented with a complete face from which features were exchanged. Secondly, while the paint package used was internal to PRO-fit, Adobe Photoshop was used for FACES.

Once a composite had been constructed, witnesses provided a 'verbal comment', a description of the likeness of their composite to the target face. They also provided a likeness rating of their composite: they were given the statement, 'The composite is an accurate likeness of the target' and accordingly rated from 1 to 5 (1 = strongly disagree / 2 = disagree / 3 = unsure / 4 = agree / 5 = strongly agree). They were then informed that another composite would be constructed using a second system. This part used the same procedure as above, with an initial composite assembled by the operator, then feature selection and artistic enhancement carried out under the direction of the witness. The entire procedure took about 1.5 – 2 hours per person.

2.2. Composite evaluation

Four tasks were used to evaluate the twenty composites (i.e. 10 from PRO-fit and 10 from FACES). The most forensically interesting was spontaneous or uncued naming and required one set of participants to identify the celebrity faces depicted in each composite. Since they would be unable to identify a composite if they did not know the person depicted, they were then asked to name the original target photographs. As correct naming levels are generally low when composites are constructed after two days, we also administered a novel 'cued' naming task, whereby participants were required to name the composites again after having seen the target photographs. It was expected that knowledge of the identities might trigger a correct response for those composites with good enough likenesses and thereby elevate performance.

Two standard supplementary tasks were also administered. The first was composite sorting, which involved fresh participants matching the composites to their target photographs. This provides a measure of the quality of features in the composites, since

participants typically compare features when carrying out the task (Frowd et al. 2005b). Also administered was a likeness rating task, whereby a further group of participants rated the subjective quality of the composites in the presence of a target photograph.

As participants inspected all 20 composites in each of the tasks, composite system was a within-subjects factor (FACES / PRO-fit).

2.2.1. Participants. Separate groups of twelve participants were recruited for each task. Composite naming was carried out by students at Stirling University and there were six male and six female volunteers, ranging from 19 to 34 years (M = 23.3 years, SD = 5.6). Composite sorting comprised of staff and students at Stirling. There were six males and six females, receiving £2, and aged from 19 to 65 years (M = 33.9 years, SD = 15.5). The composite likeness ratings were given by undergraduates at Arizona State University. These were volunteers, eight females and four males, from 18 to 25 years (M = 20.0 years, SD = 1.9).

2.2.2. Procedure. Participants were tested individually and informed that they would be evaluating a set of composites of famous faces.

Participants in the naming task were presented with the composites sequentially and asked to provide a name, where possible, for each. This procedure was repeated for the target photographs and then again for the composites. No feedback was given as to the accuracy of response given during naming.

A different group of participants completed the sorting task. They were given the pile of composites and asked to match them to the target photographs laid out on a table in front of them. They were requested to form a pile in front of each photograph, work independently of other matches and try not to make exchanges once placed on the table. They were not told anything about how many should be in each pile. A third group of participants was presented with each composite along with the target photograph and asked to provide a likeness rating (1 = poor likeness / 7 = good likeness).

All tasks were self-paced. The order of stimuli presentation was randomised for each person.

2.2.3. Results. Performance of PRO-fit and FACES composites by task is summarized in Table 1. Out of a possible 240 attempts (12 participants x 20 composites), there was only one correct name given: the PRO-fit of the actor James Redmond. This low level of 'uncued' naming was in spite of the targets being correctly named 77.5% of the time. Thus, the overall correct composite naming was 0.5% (1 / (240 * 0.775) * 100%). No inferential statistics were carried out due to low values. It should be noted that the number of incorrect names given was similar for both PRO-fit (M = 7.5%, SD = 12.2) and FACES (M = 6.7%, SD = 10.7) composites ($t_{11} = 0.22$, $p > .05$).

Higher levels were produced for cued naming (M = 5.3%). This time, PRO-fit composites elicited significantly more correct names than those from FACES (M = 9.2% vs. 2.5%, $t_{11} = 3.55$, $p = .005$) and the effect size was large, d = 1.0. Performance in the sort task yielded 31.4% overall correct and was also significantly higher for PRO-fit (M = 37.5%) than for FACES (M = 29.2%) composites ($t_{11} = 2.42$, $p < .05$); the effect size was medium, d = 0.47. Finally, while composites were poorly rated, on average 2.9 out of 7,

ratings once again favoured PRO-fit composites (M = 3.3) than those from FACES (M = 2.5, $t_{11} = 3.91$, $p < .005$), $d = 0.73$.

Table 1. Performance of PRO-fit and FACES composites.

	Uncued naming	Cued naming	Sorting	Likeness
PRO-fit	1.0 (2.9)	9.2** (7.9)	37.5* (18.6)	3.3** (1.1)
FACES	0.0 (0.0)	2.5 (4.5)	29.2 (15.1)	2.5 (1.0)

Note. Figures are percent correct, except for likeness, which are mean ratings (range 1 to 7). Numbers in columns differ significantly: * $p < .05$, ** $p <= .005$. Numbers in brackets are standard deviations.

2.2.4. Supplementary analyses. The overall mean likeness rating from witnesses themselves was 3.0 out of 5. Analysed by system, ratings were significantly higher for PRO-fit composites (M = 3.4) than for those from FACES (M = 2.5, $t_9 = 3.25$, $p = 0.01$). Therefore, the ratings collected from both witnesses and participants, while on a different rating scale, both favoured composites from PRO-fit.

An informal analysis was carried out on witnesses' verbal comments. A total of six people reported difficulty locating one or more facial features with FACES. In addition, five witnesses commented that their FACES composites suffered from what is perhaps best described as a 'holistic' deficit: their composite was female looking, had a lifeless appearance or was too young. Comments about PRO-fit composites were generally more positive, thus reflecting the higher likeness scores, although on five occasions, witnesses reported that there was something wrong with their composite but could not recall why; other comments were minor: one witness had difficulty locating an appropriate mouth, another was not satisfied with hair, and a third reported that their composite appeared too old.

2.2.5. Discussion. This experiment was a partial replication of a previous study, Frowd et al. (2005a), here involving a single composite operator and each witness constructing a composite using PRO-fit and FACES. In the Frowd et al. (2005a) study, uncued naming was very low, and did not differentiate between the systems, but performance was also equivalent in a sorting task. In the current study, all measures other than uncued naming, which was at floor, favoured PRO-fit, and with a medium (sorting) or large effect size (naming and likeness). Indeed, the superiority of PRO-fit was further reflected in the rating scores of composites from witnesses themselves.

3. Experiment 2 – PRO-fit versus Identikit 2000

Frowd et al. (2005a) found that all other composite systems tested were inferior to a sketch artist. While it cannot be denied that the possibility that the skill of this person was responsible, so might the mode of representation. It turns out that sketch artists working with witnesses tend to produce composites with less shading compared with systems such as PRO-fit and FACES that utilise photographed features, especially in areas around the forehead, cheeks, nose, and chin. This situation may arise as relatively uniform areas, such as the cheeks, are particularly hard to describe and are better left blank rather than including potentially misleading shading information. As such, sketched composites may contain less incorrect information and may be more identifiable.

One of the best known sketch-based systems is Identikit, first becoming popular in the US some 30 years ago. Like the British Photofit, it is a non-computerized method and was found to perform poorly (e.g. Laughery and Fowler 1980). Mac-A-Mug Pro is a modern computerized sketch-based system and has been shown to perform quite well when the target is present during construction (Cutler et al. 1988, Wogalter and Marwitz 1991), but composites produced from a person's memory are poorly identified (Koehn and Fisher 1997, Kovera et al. 1997). However, Identikit 2000 is a more recent and more popular sketch system than the Mac-A-Mug Pro, with a reported use in the US second only to FACES (McQuiston-Surrett et al. 2006). To the authors' knowledge, Identikit 2000 has not been the subject of a formal evaluation.

The following experiment thus compares PRO-fit with Identikit 2000. If the advantage found for sketches (Frowd et al. 2005a) is in part because of the mode of the images, then it would be expected that the more sketch-like Identikit 2000 would perform better than PRO-fit.

3.1. Composite construction

In this experiment, 10 participant-witnesses inspected a target photograph for one minute and two days later received a Cognitive Interview and constructed a composite using both PRO-fit and Identikit 2000. A second operator was trained 'in house' and then practiced equally on both Identikit 2000 and PRO-fit in order to achieve a similar level of operator skill as the operator from Experiment 1.

3.1.1. Participants. Ten undergraduates from Arizona State University were each paid $15. There were seven women and three men, aged 19 to 53 years, with a mean of 26.7 years (SD = 10.4).

3.1.2. Materials. The 10 target photographs and PRO-fit 3.1 from Experiment 1 were used, along with Identikit 2000 (marketed by Smith and Wesson, http://www.smith-wesson.com/) and Adobe Photoshop 7.0.

3.1.3. Procedure. Participant-witnesses followed the same procedure as in Experiment 1, except that PRO-fit and Identikit 2000 were the composite systems used. This time, no

participants were dismissed on the basis of being familiar with all the targets. Adobe Photoshop was used for artistic enhancement with Identikit 2000.

3.2. Composite evaluation

3.2.1. Participants. Three groups of 12 participants were drawn from the same population as Experiment 1 and given the same incentives. For naming, there were nine females and three males, aged 19 to 34 years (M = 23.5 years, SD = 5.4); for sorting there were eight females and four males, aged 17 to 48 years (M = 29.1, SD = 10.0); and for likeness ratings, there were eight females and four males, aged 18 to 24 years (M = 20.2, SD = 2.0).

3.2.2. Results. Performance of PRO-fit and Identikit 2000 composites are summarised in Table 2. Similar to Experiment 1, uncued naming was very poor, and only one correct name was elicited overall, the Identikit 2000 of Ronan Keating. Similarly, target naming was high at 77.5%, and the overall level of uncued naming was 0.5%. The incorrect composite naming was 7.5% (SD = 9.7) for PRO-fit, as before, but was higher for Identikit at 14.2% (SD = 19.3), although this difference was not reliable ($t_{11} = 1.6$, $p > .05$). Cued naming was 5% overall and varied little by system: 5.8% (SD = 6.7) for PRO-fit and 4.2% (SD = 7.9) for Identikit ($t_{11} = 1.0$, $p > .05$).

As for cued naming, system performance varied little by sorting, with PRO-fit performing at 24.2% correct and Identikit at 23.3% ($t_{11} = 0.9$, $p > .05$). While naming and sorting failed to differentiate the systems, likeness ratings favoured Identikit (M = 3.1) over PRO-fit (M = 2.6) composites ($t_{11} = 3.2$, $p < .01$), d = 0.57.

Table 2. Performance of PRO-fit and Identikit 2000 composites.

	Uncued naming	Cued naming	Sorting	Likeness
	0.0	5.8	24.2	2.6
PRO-fit	(0.0)	(6.7)	(19.3)	(0.9)
	1.0	4.2	23.3	3.1*
Identikit 2000	(2.9)	(7.9)	(18.3)	(0.8)

Note. Figures are percent correct, except for likeness, which are mean ratings (range 1 to 7). Numbers in columns differ significantly: * $p < .01$. Numbers in brackets are standard deviations.

3.2.3. Supplementary analyses. The overall mean likeness rating from witnesses was similar to Experiment 1 (M = 3.1) and was only slightly higher for Identikit (M = 3.1) than PRO-fit composites (M = 3.0, $t_9 = 0.4$, $p > .05$). The verbal comments given expressed a general dissatisfaction with the composites produced from both systems. There were a few specific remarks made, with one witness commenting on the limited options available in Identikit, another expressing dissatisfaction with the hairstyles available in PRO-fit, and a third mentioning that their PRO-fit appeared too old. On balance, the comments did not favour either system: there were two witnesses who preferred their PRO-fit composites and another two who preferred their Identikit composites, thus reflecting the equivalence in rating scores.

Two further analyses were run on the data combined across experiments to investigate whether the above results were sensible and in line with previous research. Three scores were first calculated for each participant-witness: the mean score for their two composites in cued naming, sorting, and likeness ratings. T-tests were then run by witness gender, which revealed no significant difference by cued naming ($t_{18} = 0.19$, $p > .05$), sorting ($t_{18} = 0.56$, $p > .05$), and likeness ratings ($t_{18} = 0.63$, $p > .05$). This result fits with previous research suggesting that witness gender has little influence on composite quality (e.g. Ellis et al. 1980, Frowd et al. 2005b). Next, point-biserial correlations were found to be low and non-significant between witness age and (a) cued naming ($r = -.15$), (b) sorting ($r = .35$), and (c) likeness ratings ($r = .20$). Again, these data reflect a similar story reported previously (e.g. Frowd et al. 2005b).

3.2.4. Discussion. In this experiment, participant-witnesses constructed composites using both PRO-fit and Identikit 2000. As before, uncued naming accuracy was very low. Contrary to expectations, both cued naming and sorting suggested that the systems were equivalent, although likeness ratings favoured Identikit 2000. Thus, while composites from Identikit 2000 may appear better than those from PRO-fit, the more objective measures suggest that the two systems were very similar.

4. General Discussion

The current study recruited laboratory witnesses to construct facial composites under conditions which were broadly similar to those used by law enforcement in the UK and US. In Experiment 1, PRO-fit composites were of better quality compared with those produced from FACES both on rating scores given by witnesses and on three of the four measures; only in uncued naming were the systems not differentiated, presumably due the low number of names elicited. Experiment 2 produced the same low level of uncued naming. However, while Identikit 2000 composites were rated better by third persons than those from PRO-fit, in all other measures the systems performed equivalently. Across the two experiments, as the ‘reference’ system used here, PRO-fit, performs the same as the other UK system, E-FIT (e.g. Brace et al. 2000, Davies et al. 2000, Frowd et al. 2004, 2005a, 2005b). The data suggest that the UK computerized systems generally perform the same as Identikit 2000, but better than FACES.

Frowd et al. (2005a) could not differentiate between the composites produced from PRO-fit and FACES, but the more powerful methodology employed here has allowed reliable differences to emerge (such a design might also be applicable to future studies). While composites produced from both systems were generally poor by spontaneous naming, FACES 3.0 would appear to be less effective on the other measures used. The medium to large effect sizes would also suggest that the difference between the systems is of practical value. One particularly worrying problem with FACES is that witnesses may select features in isolation to a whole face, a procedure which is known to be non-optimal (e.g. Davies and Christie 1982, Tanaka and Farah 1993). There were also a sizeable proportion of participant-witnesses who expressed a general dissatisfaction with their composite, commenting for example on the inappropriateness of the age or gender of the face (referred to as a ‘holistic’ deficit above). It would appear that this issue has emerged

due to the combined feature database, and thus the manufacturers of FACES (IQ Biometrix) might be well advised to partition their database, as done elsewhere (e.g. E-FIT, PRO-fit, Identikit 2000). Since conducting this research, IQ Biometrix has produced a new version of FACES (i.e. version 4.0). However, the authors believe that the modifications made do not improve upon the general deficits outlined here.

Identikit 2000 performed similarly to PRO-fit on two of the more objective tests of composite quality: viz. cued naming and sorting. That Identikit 2000 was rated better by third persons does suggest that caution should be applied when evaluating a composite system by appearance alone. It is possible that the more sparse representation of Identikit composites contain fewer inaccuracies, and therefore appear better, but are in fact no better than other systems on a more objective paradigm. One motivation for using Identikit 2000 was that its sketch-like composites might parallel the sketch artist in Frowd et al. (2005a) and similarly give rise to a lift in performance. The present study suggests that merely shifting to a different image modality does not necessarily improve performance; there is evidence that reducing photographs of faces to line drawings tends to inhibit recognition (e.g. Davies 1983a), unless regions of the face in shadow are taken into account (Bruce, et al. 1992). There are of course other factors that may affect performance. It could be, for example, that sketched features are better, but the distribution of features within Identikit 2000 is poor – a problem that was identified with both Photofit (Davies 1983b) and Identikit (Laughery and Fowler 1980). However, witnesses in Experiment 2 did not generally comment on specific difficulties. Instead, their verbal comments reflected a general dissatisfaction: the same as the PRO-fit witnesses in Experiment 1. It is more likely that the general problem of segmenting a face from memory for the purpose of composite construction is a very difficult task irrespective of the system employed, an issue raised in the Photofit era (e.g. Davies 1983b).

It should be noted that the PRO-fits produced by the operator in the first experiment were of better quality generally than those produced in the second, and approached significance on two of the supplementary measures, sorting ($t_{10} = 1.82$, $p < .1$, $d = 0.68$) and likeness ratings ($t_{10} = 1.90$, $p < .1$, $d = 0.67$). Therefore, in spite of an attempt made to equate the training and experience of our operators, the ability to use composite systems appears to vary naturally between operators (the operators were even of similar age and educational level). Indeed, the two operators employed here promoted differences in composite quality with an effect size similar to that found between the systems, although there was no significant difference by the arguably more important naming measures, but this does nonetheless underscore the value of a single operator when comparing composite systems. It also suggests that an operator effect applies not just to Photofit, as found previously (Christie et al. 1981, Davies et al. 1983), but also to a modern system, PRO-fit.

It would appear that top-end performance has been reached for modern composite systems and at a level that is far from ideal. Such a conclusion is worrisome, especially as composites remain an important tool for law enforcement. There is perhaps a glimmer of hope if consideration is given to the process used by the sketch artist in Frowd et al. (2005a), whose composites were found to be a little better named relative to the other systems. It turns out that the artist focused more at a global level during construction rather than on features. Construction began with a consideration of the basic proportions of the face, then the features were drawn as faint outlines. Thereafter, the detail of the composite

was increased by working on groups of features. As such, the technique has been found to be holistic in nature (Davies and Little 1990), and were it to be mirrored with computerized systems, performance may be similarly enhanced. One might, for instance, blur the composite face initially, to focus on facial proportions and the general impression of the face, then allow fine tuning later using a sharper image. This provides an interesting avenue for research to pursue in the future.

To conclude, the present results suggest that when making a choice between composite systems in the US, FACES is not a system which produces the best quality composites. In contrast, Identikit 2000 would appear to be a more appropriate choice (or the British PRO-fit). However, there is mounting evidence for poor performance from these systems when used two days after seeing a target face. It would appear that US law enforcement might like to consider constructing composites earlier than two days, which should yield a more identifiable composite (Brace et al. 2000, Bruce et al. 2002, Davies et al. 2000, Frowd et al. 2004, 2005a). Alternatively, they might employ a sketch artist, for which performance appears to be a little better (Frowd et al. 2005a) and there is good customer satisfaction (McQuiston-Surrett et al. 2006), or a so-called 'third generation' (holistic) systems such as EvoFIT that are emerging with higher performance (e.g. Frowd et al. 2006a, 2007b). Law enforcement may also like to consider publishing a composite image in a new format, one that involves viewing the composite as a series of facial caricatures (a caricature is an exaggeration of the features of a face from a reference face). This format has been recently shown to dramatically improve (uncued) naming (Frowd et al. 2007c).

References

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

BRUCE, V., HANNA, E., DENCH, N., HEALEY, P. and 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. and RARITY, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. *Journal of Applied Psychology*, **87**, 894-902.

CHRISTIE, D., DAVIES, G.M., SHEPHERD, J.W. and 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. and PENROD, S. D. (1988). An empirical examination of a computerized facial composite production system. *Forensic Reports*, **1**, 207-218.

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

DAVIES, G.M. (1983b). Forensic face recall: the role of visual and verbal information. In S.M.A. Lloyd-Bostock and B.R. Clifford (Eds.). *Evaluating Witness Evidence* (pp. 103-123). (Chichester: Wiley).

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

DAVIES, G.M. and LITTLE, M. (1990). Drawing on memory: Exploring the expertise of a police artist. *Medical Science and the Law*, 30, 345-354.

DAVIES, G.M., MILNE, A. and SHEPHERD, J. (1983). Searching for operator skills in face composite reproduction. *Journal of Police Science and Administration*, **11**, 405-9.

DAVIES, G.M. and VALENTINE, T. (2006). Facial composites: Forensic utility and psychological research. In R.C.L. Lindsay, D.F. Ross, J.D. Read and M.P. Toglia (Eds.), *Handbook of Eyewitness Psychology.* (Vol. 2. pp. 59-86). (Mahwah NJ: Erlbaum).

DAVIES, G.M., VAN DER WILLIK, P. and MORRISON, L.J. (2000). Facial Composite Production: A Comparison of Mechanical and Computer-Driven Systems. *Journal of Applied Psychology*, **85**, 119-124.

ELLIS, H., DAVIES, G.M. and SHEPHERD, J. (1978). A critical examination of the photofit system for recalling faces. *Ergonomics*, **21**, 297-307.

ELLIS, H. SHEPHERD, J. and DAVIES, G.M. (1975). Use of photo-fit for recalling faces. *British Journal of Psychology*, **66**, 29-37.

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

FROWD, C.D., BRUCE, V., MCINTYRE, A. and HANCOCK, P.J.B. (2007a). The relative importance of external and internal features of facial composites. *British Journal of Psychology*, **98**, 61-77.

FROWD, C.D., BRUCE, V., MCINTYRE, A., ROSS, D., FIELDS, S., PLENDERLEITH, Y. and HANCOCK, P.J.B (2006a). Implementing holistic dimensions for a facial composite system. *Journal of Multimedia*, **1**, 42-51.

FROWD, C.D., BRUCE, V., NESS, H., THOMSON-BOGNER, C., PETERSON, J., MCINTYRE, A. and HANCOCK, P.J.B. (2007b). Parallel approaches to composite production. *Ergonomics*, **50**, 562-585.

FROWD, C.D., BRUCE, V., PLENDERLEITH, Y. and HANCOCK, P.J.B. (2006b). Improving target identification using pairs of composite faces constructed by the same person. *IEE Conference on Crime and Security,* (pp. 386-395). London: IET.

FROWD, C.D., BRUCE, V., ROSS, D., MCINTYRE, A. and HANCOCK, P.J.B. (2007c). An application of caricature: how to improve the recognition of facial composites. *Visual Cognition*, **15**, 1-31.

FROWD, C.D., CARSON, D., NESS, H., MCQUISTON, D., RICHARDSON, J., BALDWIN, H. and HANCOCK, P.J.B. (2005a). Contemporary Composite Techniques: the impact of a forensically-relevant target delay. *Legal and Criminological Psychology*, **10**, 63-81.

FROWD, C.D., CARSON, D., NESS, H., RICHARDSON, J., MORRISON, L., MCLANAGHAN, S. and HANCOCK, P.J.B. (2005b). A forensically valid comparison of facial composite systems. *Psychology, Crime and Law*, **11**, 33-52.

FROWD, C.D., HANCOCK, P.J.B. and CARSON, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Psychology (TAP)*, **1**, 1-21.

FROWD, C.D., MCQUISTON-SURRETT, D., KIRKLAND, I. and HANCOCK, P.J.B. (2005c). The process of facial composite production. In A. Czerederecka, T. Jaskiewicz-Obydzinska, R. Roesch and J. Wojcikiewicz (Eds.). *Forensic Psychology and Law* (pp. 140-152). (Krakow: Institute of Forensic Research Publishers).

GEISELMAN, R.E., FISHER, R.P., MACKINNON, D.P. and HOLLAND, H.L. (1986). Eyewitness memory enhancement with the cognitive interview. *American Journal of Psychology*, **99**, 385-401.

GIBLING, F. and 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.

KOEHN, C.E. and 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. and THILL, D.L. (1997). Identification of computer generated facial composites, *Journal of Applied Psychology,* **82**, 235-246.

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

MCQUISTON-SURRETT, D., TOPP, L. D. and MALPASS, R. S. (2006). Use of facial composite systems in U.S. law enforcement agencies. *Psychology, Crime and Law,* **12**, 505–517.

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

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