British Journal of Psychology (2007), 98, 61–77
© 2007 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

The relative importance of external and internal features of facial composites

Charlie Frowd[1]*, Vicki Bruce[2], Alex McIntyre[1] and Peter Hancock[1]

[1]Department of Psychology, University of Stirling, UK
[2]College of Humanities and Social Science, University of Edinburgh, UK

Three experiments are reported that compare the quality of external with internal regions within a set of facial composites using two matching-type tasks. Composites are constructed with the aim of triggering recognition from people familiar with the targets, and past research suggests internal face features dominate representations of familiar faces in memory. However the experiments reported here show that the internal regions of composites are very poorly matched against the faces they purport to represent, while external feature regions alone were matched almost as well as complete composites. In Experiments 1 and 2 the composites used were constructed by participant-witnesses who were unfamiliar with the targets and therefore were predicted to demonstrate a bias towards the external parts of a face. In Experiment 3 we compared witnesses who were familiar or unfamiliar with the target items, but for both groups the external features were much better reproduced in the composites, suggesting it is the process of composite construction itself which is responsible for the poverty of the internal features. Practical implications of these results are discussed.

The perception of a face varies according to personal experience. At one extreme, a face may have only been seen once, often referred to as an 'unfamiliar' face, and the resulting memory is fragile and strongly modulated by image-specific factors such as lighting (Bruce, 1982; Hill & Bruce, 1996), viewing angle (Bruck, Cavanagh, & Ceci, 1991; Davies & Milne, 1982; Hill & Bruce, 1996), facial expression (Bruce, 1982; Davies & Milne, 1982) and background (Memon & Bruce, 1983). At the other extreme, given many encounters, identification becomes more robust and is largely invariant of such factors (e.g. Bruce *et al.*, 1999; Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999; Young, Hay, McWeeny, Flude, & Ellis, 1985).

In general, familiar faces enjoy high recognition (e.g. Bruce, 1988; Bruce *et al.*, 2001), even under difficult conditions such as poor quality video (Burton *et al.*, 1999), highly pixelated images (e.g. Bruce & Young, 1998) or brief presentation (Lee & Perrett, 1997). The misidentification of familiar faces is very low (e.g. Young, Hay, & Ellis, 1985).

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

DOI:10.1348/000712606X104481

In contrast, under the influence of these image-specific factors, unfamiliar faces tend to be identified less well than familiar faces, but misidentified much more often, even under very favourable conditions such as high quality media (e.g. Henderson, Bruce, & Burton, 2001). Indeed, misidentification, through unfamiliar face processing, is believed to be the principal cause of wrongful conviction (Rattner, 1988).

The relative importance or salience of different areas of the face also appears to change with familiarity. Ellis, Shepherd, and Davies (1979) found that familiar faces were recognized more accurately from their internal features – the region including the eyes, brows, nose and mouth – than from their external facial features – head shape, hair and ears – but for unfamiliar faces, the external features dominated. Similarly, Young *et al.* (1985) found faster reaction times and fewer errors made for the internal features of familiar faces using a face-matching paradigm. In general, an advantage for the inner face of familiar items emerges with adulthood (Campbell *et al.*, 1999) and may be observed following relatively few encounters (O'Donnell & Bruce, 2001). For unfamiliar faces, the external parts of a face tend to have a more salient role in face perception than their internal counterparts (Bruce *et al.*, 1999; de Haan & Hay, 1986; Ellis *et al.*, 1979; Hancock *et al.*, 2000; Young *et al.*, 1985).

One applied area of psychology that involves both familiar and unfamiliar face processing is the production and recognition of facial composites. These are pictures of suspects to a crime produced by witnesses (or victims) with the help of computer software or a sketch artist. In general, witnesses see a suspect for a short time – maybe only for a few seconds – and so composite construction engages unfamiliar face perception, although the goal is to promote identification by people familiar with the target when a composite is published. While composites remain an important tool for the apprehension of suspects, research has suggested that they are of poor quality and generally only named around 20% of the time, at best, when witnesses construct them soon after seeing a target face (Bruce, Ness, Hancock, Newman, & Rarity, 2002; Brace, Pike, & Kemp, 2000; Davies, van der Willik, & Morrison, 2000; Frowd *et al.*, in press; Frowd, Carson, Ness, Richardson, *et al.*, 2005; Frowd, Hancock, & Carson, 2004). Unfortunately, composite quality appears to be even worse when construction occurs 2 days after inspecting a target face, as tends to occur in real life (Frowd *et al.*, in press; Frowd, Carson, Ness, McQuiston, *et al.*, 2005; Koehn & Fisher, 1997).

As unfamiliar faces are remembered and processed generally better by their external features, and witnesses typically engage in this type of processing, it was predicted that the external features of facial composites should be better constructed than their internal ones. However, given the relative importance of the internal features of a face for identifying a person (e.g. Bruce *et al.*, 1999; Ellis *et al.*, 1979; Young *et al.*, 1985), it is possible that composites are not well named due to their poor representation of internal features – the parts of a face most critical for identifying familiar people.

The focus of the current paper is twofold. In the first two of three experiments, we investigate the relative recognizability of the internal and external features of composites produced using procedures that mirror some aspects of real life: for example, from participant-witnesses who were unfamiliar with a target face and waited a couple of days prior to construction of composites from memory. Our prediction is that the quality of the internal composite features should be worse than the quality of the external composite features, as indicated by two matching-type tasks. This part also asks whether there is any useful information in general conveyed by the internal features. In Experiment 3, we went on to investigate the reasons why internal features are poorly depicted in the composites, investigating whether the problem concerns the

unfamiliarity of the target faces or some inherent difficulty in constructing internal vs. external features.

INTRODUCTION TO EXPERIMENTS 1 AND 2

The materials used for the first two experiments were derived from a set of composites and target photographs used by Frowd, Carson, Ness, McQuiston, *et al.* (2005). The reader is referred to this paper for full details of the procedures used to construct these composites. In brief, participant-witnesses inspected a photograph of an unknown male, described his face and constructed a composite 2 days later working with an operator using one of five different composite systems. The methods included three popular computerized systems – E-FIT (UK), PRO-fit (UK) and FACES (US) – that required witnesses to assemble a face by selecting facial features from a predefined set of parts (hair, face shape, eyes, noses, mouth, etc); a UK police sketch artist, who drew out the composite face by hand; and an early version of EvoFIT, a UK 'holistic' system in development (Frowd *et al.*, 2004), where a composite was 'evolved' through the repeated selection and 'breeding' of whole faces.

Frowd, Carson, Ness, McQuiston, *et al.*'s (2005) study used 10 target faces. These were of celebrities generally unknown to people aged over 30, who were the witnesses in the study, but familiar to participants later in the experiment, who were in their early twenties and carried out the main evaluation by naming. Further details of the targets may be found in Materials, Experiment 1 (below). The study also provided a check to verify that the targets were not recognized by participant-witnesses and thus the composites were constructed of unfamiliar faces.

After it had been established that the target was unfamiliar, participant-witnesses first inspected a photograph of a celebrity face for 1 minute in the knowledge that a composite would later be required (i.e. intentional learning was employed). Two days later, each participant-witness constructed a single composite using procedures that followed real witnesses as far as possible. These procedures included: the use of experimenters who were experienced in assisting witnesses construct a composite (referred to here as composite operators for computerized systems); a cognitive interview, to assist recall of the target face; open-ended construction sessions, to promote good performance; and software painting tools, whose use may improve the appearance of a face (e.g. Christie, Davies, Shepherd, & Ellis, 1981; Davies, Milne, & Shepherd, 1983; Frowd, Carson, Ness, Richardson, *et al.*, 2005; Geiselman, Fisher, MacKinnon, & Holland, 1986; Gibling & Bennett, 1994). Composites took about an hour to construct using the E-FIT, PRO-fit and FACES systems, and about 2 hours for Sketch and EvoFIT.

Composites of each of the 10 celebrity targets were constructed using each of the five systems to produce 50 composites in total. Evaluation of the quality of the resulting composites involved further participants and three separate tasks: naming, where participants attempted to name the composites; sorting, where participants attempted to match the composites to the target photographs; and a line-up task, where participants attempted to identify the target photographs from a six-item photo spread. The results from these tasks found that while Sketch was the best system by both naming and sorting, E-FIT was better than all others in line-ups, although not significantly better than Sketch. It was also found that the composites were generally of poor quality, being named only about 3% of the time overall, and being matched

correctly only about 40% of the time in the sorting and line-up tasks, in spite of using the same photograph of the target as that used for their construction.

The current study used 40 of the composites constructed in Frowd, Carson, Ness, McQuiston, *et al.* (2005) – those produced from the UK systems: E-FIT, PRO-fit, Sketch and EvoFIT (The US FACES composites were omitted to create a more manageable set.)

EXPERIMENT 1

A sorting task was used first to explore the relative quality of the external and internal features of the composites. It is a standard task used to evaluate composites (e.g. Davies *et al.*, 2000; Frowd, Carson, Ness, McQuiston, *et al.*, 2005; Frowd, Carson, Ness, Richardson, *et al.*, 2005; Wogalter & Marwitz, 1991) and involves participants matching composites to their target photographs. It provides an indication of feature quality as participants tend to compare facial features between composites and targets (Frowd, Carson, Ness, Richardson, *et al.*, 2005).

The work also explored whether the participants recruited here would be influenced by how familiar they were with the target faces. It is conceivable, for example, that being more familiar with a target face would result in an improved ability to sort the internal parts of a composite and vice versa. This notion was investigated by recruiting participants from a wide age range to permit a natural variation in target familiarity – recall that the targets were selected to be generally unknown to people over the age of 30 – and allow an analysis by participant familiarity.

In this experiment, participants inspected all 40 composites in one of three conditions – complete (veridical) composites, internal composite features or external composite features – and therefore composite type (complete/internal/external) and participant familiarity (low/high) were between-subjects factors, but composite system (E-FIT/PRO-fit/Sketch/EvoFIT) was a within-subjects factor. We anticipated that the external composite features would be sorted better than the internal composite features, and the complete composites would be best of all. In addition, the internal composite features were expected to be sorted better when participant familiarity with the targets was high rather than low.

Method

Participants

Thirty staff and students from Stirling University were paid £2 to sort the composites. There were 15 males and 15 females, aged 18 to 60 years, with a mean age of 29.2 years ($SD = 11.7$).

Materials

The stimuli were the target photographs and composites produced from Frowd, Carson, Ness, McQuiston, *et al.* (2005). The targets were photographs of 10 celebrities of actors (Ben Affleck, Matt Damon, Jeremy Edwards, Joshua Jackson, Philip Olivier and James Redmond) and pop singers (Kian Egan, Mark Feehily, Ronan Keating and Ian 'H' Watkins). Each face was clean shaven as far as possible, and spectacles were avoided. Faces were printed in colour on a good quality printer with dimensions of approximately 6 cm (wide) × 8 cm (high).

Three sets of composites were used. These were the set of 40 from Frowd *et al.* which had been constructed from E-FIT, PRO-fit, Sketch and EvoFIT, referred to as 'complete' composites, plus two additional preparations: a set containing the internal features and another set containing the external features. The internal composite features were prepared in Adobe Photoshop by highlighting the internal facial features of these composites in the shape of an oval and then truncating just above the eyebrows, to omit the forehead that sometimes contained hair. As can be seen in Figure 1, the external composite features were produced by simply covering the previously defined internal features with a uniform grey mask.

Procedure

Participants were tested individually and randomly assigned, with equal sampling, to one of three composite sets (complete/internal features/external features) that contained 40 composites. They were told that they would be evaluating composites of famous faces by matching them to their celebrity targets. The target photographs, which were always complete faces, were then placed on the table in front of each person and the relevant pile of 40 composites given. Participants were asked to work through the set sequentially by placing each composite in front of a celebrity face, in their own time, but to try not to make exchanges once placed on the table. Using this procedure, the composites were thus sorted. The order of presentation was randomized across the set to mix the composites from the different systems (i.e. only one block was used). The presentation order of targets was also randomized for each person.

Results and discussion

Composite performance is shown in Table 1. Overall, whole composites and those of external features were sorted similarly, at approximately 33% correct, and were both appreciably higher than composites of internal features, at 19.5%; the effect size was medium to high: whole vs. internal ($d = 0.64$); external vs. internal ($d = 0.63$). E-FIT was the best system for both complete and external features composites. There was little difference by system for composites of internal features.

The sorting scores for complete, external and internal feature composites were each partitioned equally by participant age to permit an analysis by participant familiarity with the targets. This provided a mean participant age of less than 22 years for scores in

Figure 1. Example stimuli of the UK pop singer Ian 'H' Watkins. Participants inspected complete composites (left), internal composite features (centre) or external composite features (right).

Table 1. Percent correct performance in the composite sorting task

	Complete	External	Internal	Mean
E-FIT	41 (20.2)	47 (20.6)	23 (17.0)	37.0 (21.4)
PRO-fit	28 (22.5)	22 (25.3)	20 (15.6)	23.3 (21.1)
Sketch	34 (22.7)	30 (21.6)	17 (17.7)	27.0 (21.4)
EvoFIT	29 (19.7)	32 (24.4)	18 (17.5)	26.3 (20.9)
Mean	33.0 (21.4)	32.8 (24.0)	19.5 (16.5)	28.4 (21.5)

Note. The data reveal a consistent bias towards the external facial features. Figures in brackets are standard deviations.

the complete, external and internal features composites for the high familiarity condition, and over 30 years for scores in the low familiarity condition. It was found that the mean sorting scores did not change by familiarity for complete ($M = 33.0\%$) and internal features ($M = 19.5\%$) composites, and were only slightly different for participants in the high ($M = 29.0\%$) and low ($M = 34.0\%$) familiarity condition for external features.

The participant accuracy scores were subjected to a three factor mixed analysis of variance (ANOVA). This produced both a main effect of composite type ($F(2, 24) = 7.50$, $p < .005$) and system ($F(3, 72) = 6.35$, $p = .001$), but participant familiarity ($F(1, 24) = 0.07$, $p > .1$) and all the interactions did not approach significance, ($F < 1.5$). Simple contrasts suggested that the quality of complete and external features composites did not differ significantly ($p > .1$) and both were of better quality than those of internal features ($p < .05$); by system, E-FITs were marginally better ($p < .1$).

In line with expectation then, the sorting task revealed that composites of external features were of better quality than those of internal features for all systems tested. The internal composite features were sorted poorly, at 19.5% correct, a value close to, although significantly greater than, the 10% correct expected by chance ($t(39) = 3.65$, $p < .001$, by items). Given the evidence regarding the importance of internal features for the identification of familiar faces, and their ineffectiveness as measured by sorting, it is perhaps not surprising that these composites attract poor naming (Frowd, Carson, Ness, McQuiston, *et al.*, 2005).

It was also found that the external composite features were sorted as accurately as the complete composites, and therefore it is likely that the internal features were of little value when presented with hair, ears, etc. In support of this notion, point-biserial correlations were carried out on the mean sorting scores (items) between the complete, internal and external composite features. The only significant correlation was between the complete and external features group ($r(38) = 0.38, p < .05$), thus highlighting the role played by the external features in the processing of complete composites.

The participants in this experiment presented a wide age range and were therefore likely to vary in their familiarity with the target faces. It was anticipated that higher familiarity with the target faces would promote better sorting scores for internal feature composites. One might also expect that increasing familiarity would result in a greater dependence on the internal features when sorting complete composites and promote a decrease in accuracy (as composites of internal features tend to be sorted worse than those of external features). However, the experiment provided no evidence that target familiarity influenced performance in this task and therefore the differences in quality

found between external and internal features would appear to be solely a function of composite construction.

In the following experiment, we attempted to replicate the 'external features advantage' using a second method of evaluation, a photo line-up task.

EXPERIMENT 2

This experiment used a photo array to compare the quality of the external and internal composite features of Experiment 1. While more difficult to set up than a sorting task, as a number of distracter faces (foils) must be located for each target face, the photo array task (or line-up) may function more as a measure of identification, such as a police line-up, given that the arrays contain foils and targets deliberately selected to be similar to each other. A version of this paradigm with an array of five distracters plus target was employed, as used previously (e.g. Bruce *et al.*, 2002; Frowd, Carson, Ness, McQuiston, *et al.*, 2005; Koehn & Fisher, 1997).

Experiment 1 found that the internal composite features were matched at near chance levels. This time, an attempt was made to elevate performance by including an additional set of 'easy' arrays (see Materials). The design was between-subjects for array type (easy/hard) and, as in Experiment 1, composite type was between-subjects (internal/external features), while composite system (E-FIT/PRO-fit/Sketch/EvoFIT) was within-subjects.

The previous experiment also found, using a wide participant age range, that a participant's level of familiarity did not significantly influence performance on a composite matching task. Given this result, and for convenience, the current experiment employed a greater proportion of younger participants (i.e. there were more participants who were familiar with the target faces).

Method

Participants

Forty-eight undergraduates at Stirling University volunteered. There were 21 males and 27 females and their age ranged from 18 to 31 years with a mean age of 21.4 years ($SD = 2.3$).

Materials

The materials consisted of the 40 external composite features and the 40 internal composite features of Experiment 1, plus two sets of photo line-ups (complete faces). While the 'hard' line-ups were those used by Frowd, Carson, Ness, McQuiston, *et al.* (2005), and contained celebrities who appeared visually similar to each target, the 'easy' ones were modified to improve performance by creating arrays more different to each target. In practice, this involved identifying the two items in each array which were the most confusable to the target, from Frowd *et al.*'s data, and then randomly exchanging them between arrays.

In general, the foils used for the line-ups were depicted in a front view, without spectacles, and were either clean shaven or had minor facial stubble. They were printed in monochrome using a good quality printer at a size of approximately 6 cm (wide) × 8 cm (high) to enable the face arrays (target face plus five foils) to appear on one side of A4 paper.

Procedure

Participants were tested individually and assigned randomly, with equal sampling, to one of four testing books (easy[array]-internal[features]/easy-external/hard-internal/hard-external). They were told that they would be evaluating a set of composites of famous faces by picking out the targets from a photo line-up. Participants were then presented with each composite sequentially, along with the associated array, and selected a celebrity face in their own time. As in Experiment 1, the order of presentation was randomized across each booklet for each person.

Results and discussion

Overall, as can be seen in Table 2, composites of external features ($M = 41.6\%$) were identified much better than those of internal features ($M = 28.4\%$); the effect size was also large ($d = 0.90$). This result was consistent across array type, although more pronounced for easy arrays. There was little difference in identification accuracy between easy ($M = 36.6\%$) and hard ($M = 33.4\%$) arrays. By system, E-FITs ($M = 40.4\%$) performed overall much the same as Sketch ($M = 41.7\%$), and both were better than PRO-fit ($M = 29.6\%$) and EvoFIT ($M = 28.3\%$).

Table 2. Percent correct performance in the line-up task by composite type (external features/internal features) and array type (hard/easy)

	External features			Internal features			
System	Easy array	Hard array	Mean	Hard array	Easy array	Mean	Mean
E-FIT	53.3 (20.9)	48.3 (20.7)	50.8 (20.8)	30.0 (18.5)	30.0 (15.3)	30.0 (16.9)	40.4 (18.9)
PRO-fit	43.3 (27.4)	34.2 (22.4)	38.8 (24.9)	24.2 (19.4)	16.7 (16.7)	24.0 (18.1)	29.6 (23.3)
Sketch	46.7 (26.1)	36.7 (28.7)	41.7 (27.4)	44.2 (17.6)	39.2 (23.3)	41.7 (20.5)	41.7 (23.6)
EvoFIT	42.5 (24.7)	27.5 (20.1)	35.0 (22.4)	22.5 (18.9)	20.8 (18.5)	21.7 (18.7)	28.3 (21.7)
Mean	46.5 (24.3)	36.7 (23.6)	41.6 (23.9)	30.2 (19.9)	26.7 (20.0)	28.4 (18.6)	35.0 (21.9)

Note. Composites of external features ($M = 41.6\%$, $SD = 24.3$) were identified better than those of internal features ($M = 28.4\%$, $SD = 19.9$). Figures in brackets are standard deviations.

Participant scores were analysed by a three factor mixed ANOVA. This was significant for composite type ($F(1, 44) = 20.01, p < .001$), again confirming an overall advantage for external features, and system ($F(3, 132) = 12.85, p < .001$). Array type turned out not to be significant ($F(1, 44) = 1.13, p > .1$), contrary to expectation, although it interacted with composite type ($F(1, 44) = 5.16, p < .05$) as the external features advantage only applied to easy arrays ($p < .001$). Composite type also interacted with system ($F(3, 132) = 5.64, p = .001$) as there was an external feature advantage for all systems ($p < .005$) apart from Sketch ($p > .1$). No other interactions were significant ($F < 1$). For simplicity, the detailed main effects and interactions by system are not reported, although Sketch was better than all others for internal features ($p < .01$) and E-FIT showed weak evidence of being best for external features ($p < .1$).

In summary, the line-up data found a substantial advantage for external over internal composite features for all systems tested except Sketch, thus supporting the general finding of Experiment 1. The advantage of external over internal features was significant only for the easy arrays, suggesting that the changes made to produce the easy arrays

mainly affected the external features. As the most misidentified array members in Frowd, Carson, Ness, McQuiston, *et al.* (2005) were exchanged, this suggests that performance on their line-ups were also influenced by external features. Frowd *et al.* were at pains to explain why E-FIT composites were best in line-ups, although not in any other task, and our data suggest that this was due to (1) E-FIT producing composites with somewhat better quality external features and (2) line-ups being rather sensitive to external features, as suggested elsewhere with arrays and unfamiliar faces (e.g. Bruce *et al.*, 1999). It is perhaps worth mentioning that while this experiment did not include complete composites, the mean score from Frowd *et al.*'s array task was 44.5% (for E-FIT, PRO-fit, Sketch and EvoFIT composites, Table 2) and was only slightly higher than composites of external features evaluated here using hard arrays (36.7%), thus further emphasizing the important role of the external features.

For the internal composite features, the overall performance was much the same as Experiment 1 in which chance is taken into account (sorting: $M = 19.5\%$, chance $= 10\%$; line-ups: $M = 28\%$, chance $= 16.7\%$) and therefore supports the previous finding that the inner part of these composites is of poor quality. By system, sketches had consistently the best set of internal features, a result that would appear to resonate with Frowd *et al.*'s naming data, where they were named at about twice the rate of the other systems (Frowd *et al.*'s sorting data also favoured sketches).

EXPERIMENT 3

The above data suggest that the external composite features are of better quality than their internal counterparts. Why should this be? We proposed earlier that this is a probable consequence of witnesses receiving a limited exposure to a target face and engaging in unfamiliar face processing to produce a composite. However, the previous experiments do not test this hypothesis directly and our results could simply be due to a general inability of witnesses to construct the internal parts of a composite face. We test this possibility here by manipulating a witness's familiarity with a target face prior to composite construction. For convenience, a single composite system, PRO-fit, was utilized, as opposed to the four different systems in Experiments 1 and 2. PRO-fit is a typical feature-based software program in general police use in the UK, and has been found to perform equivalently to the other UK variant, E-FIT (Frowd, Carson, Ness, McQuiston, *et al.*, 2005; Frowd, Carson, Ness, Richardson, *et al.*, 2005). If our original theory is correct, then being familiar with a target face should promote composites with internal features whose quality is better than their external counterparts. Alternatively, if the problem is one of a general inability to construct composites, then the internal–external feature difference should not vary with target familiarity.

Such an investigation requires both the construction and evaluation of a set of composites. Participant-witnesses show large individual differences in composite quality, so to achieve good experimental power, we adopted a within-subjects design with each person constructing two composites, one with a familiar target and one with an unfamiliar target. This was achieved using target faces and participant-witnesses drawn from two university departments, thus allowing targets to be selected with varying familiarity. This design, used previously in this area (Davies *et al.*, 2000), involved a pre-test phase where participant-witnesses rated their familiarity with the target faces to enable appropriate target assignment. Davies *et al.* found that target familiarity did not influence composite production from a system in current police use, except when participant-witnesses had a target face in front of then during

construction. However, their participants constructed four composites in the course of an hour, and so may have been rushed and produced non-optimal likenesses. In the current study, construction will follow a more ecologically valid procedure (e.g. open-ended construction sessions). The resulting composites will be evaluated by naming as well as a complete/internal/external feature sort.

Composite construction

Participant-witnesses were first given a pre-test session to indicate familiarity with the targets and to assign them to an appropriate pair of target faces. In a separate session, they constructed two composites of these faces. The design employed an equal number of participant-witnesses from Psychology and Computing Science and was fully counterbalanced for target familiarity (low/high) and construction order (first/second). Familiarity of the target face at construction was a within-subjects factor.

Method

Participants

Eight staff and students each from Computing Science and Psychology were paid £10 to be participant-witnesses. Their age ranged from 21 to 61 years with a mean age of 32.6 years ($SD = 11.3$).

Materials

The target faces were colour photographs of four male members of staff in Computing Science and four male members of staff in Psychology at the University of Stirling. Each person was clean shaven, without spectacles and photographed in a front face pose and a neutral expression. Each face was printed in colour using a high quality printer with dimensions of approximately 10 cm wide × 13 cm high on a single sheet of A4 for presentation to participant-witnesses, and 6 cm × 8 cm for the evaluation stage. Each participant-witness saw two faces, one drawn from the Psychology set and one drawn from the Computing Science set. The set of target faces was arranged as four pairs (one target from each department), and each pair was used with four participant witnesses, with the allocation of participant-witness to pair based on assessment of prior familiarity, as described below.

Procedure

Participants were tested individually throughout. In a pre-test session, participants rated the familiarity of the target faces (1 = very unfamiliar/5 = very familiar). To do this, each of the eight faces was presented sequentially, in a random order, and as quickly as possible to limit exposure of the faces. No mention was made that the faces were the stimuli to be used later. The rating exercise was carried out typically 1–2 days before the composites were constructed.

Participants were then assigned to one of the four pairs of target faces. For each person, we identified the pairs of target faces that were rated at the extremes of familiarity and randomly assigned one of these pairs to that person, with the constraint that each target pair could only be constructed twice for participants in Computing Science (order of presentation counterbalanced) and twice for participants in Psychology (also counterbalanced).

In a separate session, the composites were constructed. Participants were assisted by a composite operator, a person experienced in composite construction for several years. Participants were given an envelope containing the first target photograph and looked at it for 30 seconds; this being carried out in the knowledge that a composite would be required. Afterwards, they were told that procedures would be followed which were similar to those used by real witnesses and an overview of these were given.

Participant-witnesses freely described the person's face in their own time and with minimal interruption from the operator. Next, the description given for each feature was repeated and they were given the opportunity to recall further. The operator then introduced the PRO-fit system, and demonstrated how facial features could be selected, positioned and resized. It was explained that as there were too many features available, the verbal description would be used to select a manageable set (normally about two dozen features). It was also mentioned that the set of available features was not exhaustive and so it was sometimes necessary to improve the likeness using a paint program within PRO-fit.

A composite was therefore constructed using this procedure. The operator used the verbal description to identify a subset of features within PRO-fit and presented each participant-witness with an initial composite: a face containing the features that best matched the description. Participant-witnesses opted to work on features of their choice, which was normally hair and face shape first, and the operator switched examples in and out of the face, with resizing and positioning, to give the best likeness. The paint program was used if required. When complete, the composite was saved to disk and participants were given a short break (normally 5–10 minutes). Following this, they inspected the second target photograph as before and used the same procedure to obtain a verbal description and construct a composite of this face. The entire procedure lasted about 1.5 to 2 hours.

In summary, the study produced a total of 32 composites, from the eight Psychology and the eight Computing Science participant-witnesses, half of which were made with a familiar target and half of which made with an unfamiliar target (order counter-balanced).

Composite evaluation

The initial evaluation of the composites followed the design of Experiment 1, with participants matching internal, external or complete composites to their target photographs. As participants inspected all 32 composites of one type (complete/external/internal), the design was between-subjects for composite type, as before, but within-subjects for target familiarity (low/high). The set of complete composites were also evaluated by naming in order to assess the general effect of familiarity on composite production.

Method

Participants

There were 30 female and 24 male participants who volunteered for the sorting task. These were aged from 16 to 56 years, with a mean age of 26.9 years ($SD = 9.6$), and were drawn from visitors at the Glasgow Science Centre and students at Stirling University. These participants were in neither Computing Science nor Psychology departments and were therefore unlikely to be familiar with the target faces. A further

15 male and 11 female staff and students volunteered to name the composites, and were selected as being likely to know the targets. These participants were aged from 20 to 54 years, with a mean age of 33.4 years ($SD = 8.8$), and were sampled equally from Computing Science and Psychology at Stirling.

Materials

As in Experiment 1, two additional sets of composites were prepared in Adobe Photoshop, one each for external and internal features.

Procedure

For the sorting task, the same procedure as Experiment 1 was employed – that is, with participants randomly assigned, with equal sampling, to three testing booklets – except that participants were not told that the composites were of famous faces.

For the naming task, participants were also tested individually and told that they would be evaluating a set of composites. However, it was explained that some of the composites were of familiar staff in their department and that participants should attempt to name them. Complete composites were used throughout. Thus, the 32 composites were presented sequentially and participants attempted to provide a name where possible. Presentation was self-paced but each composite was observed for about 10 seconds on average. The order of presentation was randomized for each person.

Results and discussion

Overall, the composites were sorted to an accuracy of 57.7%. Contrary to expectation, there was little difference in sorting scores for composites constructed when the target was rated as very familiar ($M = 50.0\%$, $SD = 19.1$) or very unfamiliar ($M = 48.7\%$, $SD = 23.1$) to participant-witnesses. In contrast, as can be seen in Figure 2, composites of external features were once again sorted much better than those of internal features ($M = 53.3\%$, $SD = 13.3$ vs. $M = 32.6\%$, $SD = 17.1$), and the effect size was very large ($d = 1.4$); as before, complete composites ($M = 61.1\%$, $SD = 16.5$) were sorted similarly to those of external features ($M = 54.3\%$, $SD = 13.3$).

These data were subjected to a two factor mixed ANOVA. This suggested that there were reliable differences by composite type ($F(2, 51) = 16.1$, $p < .001$), but neither target familiarity ($F(1, 51) = 0.36$, $p > .1$) nor the interaction ($F(2, 51) = 1.5$, $p > .1$) were significant. Simple contrasts also suggested that both complete and external features composites did not differ significantly in quality ($p > .1$), but were better than those of internal features ($p < .001$), thus mirroring the results of Experiments 1 and 2. While the two-way interaction between composite type (complete/external/internal) and familiarity (low/high) did not approach significance, Figure 2 nonetheless suggests that the external over internal feature advantage is somewhat greater for unfamiliar faces ($M = 26\%$) than for familiar faces ($M = 18\%$), and an ANOVA examining just the external and internal feature composites shows a near-significant interaction with familiarity ($p < .1$). So there is a non-significant trend suggesting that unfamiliarity with a target face may contribute to an external feature advantage for composites, but clearly other factors predominate, as discussed below.

In terms of naming, the (complete) composites were correctly named overall 22.1% of the time. While composites produced with high target familiarity ($M = 23.1\%$,

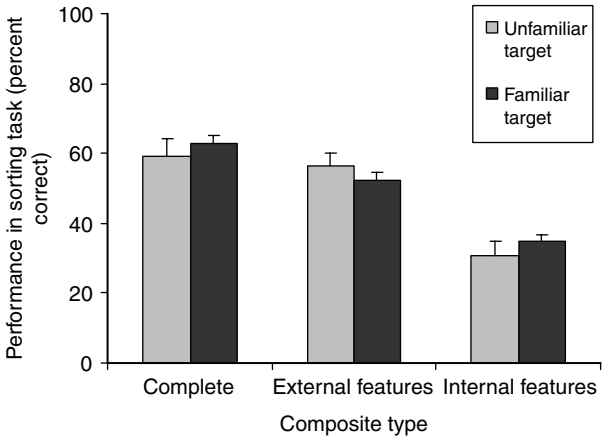

Figure 2. The effect of target familiarity on composite quality for participant-witnesses in Experiment 3. Error bars are standard errors of the means.

$SD = 20.8$) were correctly named slightly more often than when the target familiarity was low ($M = 21.2\%$, $SD = 23.7$), this difference did not approach significance ($t(25) = 0.56$, $p > .1$) and thus underscores the above result that changes in target familiarity do not appear to significantly influence composite quality.

In summary, in line with the previous two experiments, the sort task indicated that the quality of the complete and external features composites did not differ significantly, but were better than those of internal features. However, neither task suggested that this external features advantage was a consequence of low target familiarity: the quality of internal and external composite features changed little whether or not the target was very familiar. This suggests that differences in quality between internal and external features are an inherent property of composite production rather than a consequence of unfamiliar face perception.

GENERAL DISCUSSION

The facial composites investigated in Experiments 1 and 2 here were produced by participant-witnesses who did not know the identity of their target and therefore engaged in unfamiliar face processing for their construction. As such, we suggested that these composites should demonstrate a bias towards the external facial features. The data did suggest that the region including the ears, hair and face shape was of better quality than the region including the eyes, brows, nose and mouth. With the exception of sketches in the line-up task, this finding was consistent across four composite techniques and two methods of evaluation. In Experiment 3, we explored the role of target familiarity on such internal/external feature differences. While the data supported the previous finding, that complete and external composite features were of similar quality and better than those of internal features, there was no evidence that target familiarity strongly influenced composite production and thus the poor quality internal features of Experiments 1 and 2 were not a direct consequence of composite construction of an unfamiliar face.

In Experiments 1 and 2, while the external features were of better quality, participants performed only just above chance with the internal features on the tasks used. Also, in Experiments 1 and 3, the external and complete composites were of equivalent quality, suggesting that the contribution of the inner features to the appearance of a composite as a whole is minimal. Thus, given a poor set of internal features, it is not surprising that Frowd, Carson, Ness, McQuiston, *et al.* (2005) found very poor naming for these composites. Participant-witnesses who constructed them received a good exposure to their target, sufficient anyway to provide a detailed description of the face, and yet the internal features selected by witnesses appear rather ineffective at conveying identity.

Experiment 3 found that considerable exposure to a target face (a rating of 'very familiar' by participant-witnesses) did not impact upon the identifiability of complete composites or the quality of the internal features. Therefore, the advantage of the external facial features for unfamiliar faces reported generally in face perception (e.g. Bruce *et al.*, 1999; Campbell *et al.*, 1999; Young *et al.*, 1985) appears not to extend to the construction of composites. Instead, composites are naturally constructed with an inferior set of internal features; our data suggest that the internal features only increase very slightly in quality with large increases in target familiarity (from the near-significant trend in the sorting data). Only if the target is present during construction, an ecologically invalid procedure, is there any good evidence that an increase in composite quality follows an increase in target familiarity (Davies *et al.*, 2000).

Participants elsewhere in this paper, particularly those who carried out the matching-type tasks, also varied in their familiarity with the target faces. In Experiment 1, half the participants were generally familiar with the target faces and half were not; in Experiment 2, familiarity was generally good; and in Experiment 3, familiarity was generally poor. Could these varying familiarity levels influence the interpretation of the results? We believe this is unlikely, given the result from Experiment 1, which found that sorting accuracy did not significantly alter with changes in familiarity, and also from Experiments 2 and 3, where familiarity was either good or poor, but a consistent pattern of results was found (complete sorted the same as external, but better than internal). Therefore, target familiarity does not appear to play a role in the evaluation of composites using a matching-type task.

We acknowledge that the task facing participants and witnesses constructing composites is different to that generally used in face perception where the relationship between internal and external features has been studied. While face perception paradigms include recognition memory tasks, where participants are required to report whether a face has been presented previously (e.g. Burton *et al.*, 1999), and matching tasks, where faces are classified among alternatives (e.g. Bruce *et al.*, 2001) – similar to the arrays used in Experiment 2 – composite construction with real witnesses currently requires the description, selection and manipulation of individual features. Both face recognition and composite construction tasks are modulated by the type of encoding used with a target: composites are better reproduced following a feature-by-feature encoding (e.g. Wells & Hryciw, 1984), but face recognition is elevated after a holistic (trait) encoding (e.g. Shapiro & Penrod, 1986). Thus, while typical face perception tasks involve more natural or holistic (whole face) processing and allow differences in internal and external feature quality to be modulated by familiarity, composite construction from memory does not allow this, arguably due to an unnatural focus on individual features.

We have known for nearly four decades, dating back to the time when the older non-computerized approaches such as Photofit and Identikit were in regular police use, that the process of segmenting a face from memory for the purpose of description and feature selection for composite construction is at variance with the natural way faces are perceived (e.g. Davies, Shepherd, & Ellis, 1978). While modern composite systems have been enhanced considerably over the last 20 years, as evidenced by good quality composites produced when copying a photograph (e.g. Cutler, Stocklein, & Penrod, 1988; Davies *et al.*, 2000; Frowd *et al.*, in press; Koehn & Fisher, 1997), poor quality composites are still produced when working from memory (e.g. Frowd *et al*., in press; Frowd, Carson, Ness, McQuiston, *et al.*, 2005; Koehn & Fisher, 1997).

How then might facial composites be improved, especially from the computerized systems where the quality appears to be so low? One way might be to improve the internal features, and thus promote better naming for the composite as a whole. We venture that the internal features of computerized composites are poorly reproduced due to the presence and dominance of the external features during construction. Recall that witnesses tend to select the external features first, especially the hair, and then consideration is given to the internal parts. Using this procedure, the identifiability of sketches has been found to be better than composites produced by the computerized systems (Frowd, Carson, Ness, McQuiston, *et al.*, 2005; also Experiment 2 here). It turns out that while computerized systems use photographed features, sketches contain relatively little detail. As such, witnesses working with an artist may not be overwhelmed by the presence of the external features but focus on the more important, internal parts of the face, and thus produce a better quality set of internal features and a more identifiable composite (Frowd, Carson, Ness, McQuiston, *et al.*, 2005). Consequently, for a computerized system, blurring the external features at the start of construction may similarly focus attention on the internal features. Later, when the inner face has been constructed, the external part could be sharpened up to allow selection of the face shape, hair and ears. We are currently exploring this possibility.

In summary, the problem then is that composites are poor likenesses as witnesses cannot recall the features that are used by others for identification. While the external features of a familiar face may provide some clues to identity, it is the internal features that are important for triggering a name (Campbell *et al.*, 1999; Ellis *et al.*, 1979) and it is these aspects that are difficult for a witness to translate. It also appears to be the case that considerable exposure to a face, as may occur sometimes in real life, does not promote a better quality composite.

Acknowledgements

The work was funded by a grant from the Engineering and Physical Sciences Research Council, EP/C522893/1(P). The authors would like to thank the Glasgow Science Centre for allowing us to collect data for Experiment 3, plus some insightful comments from two anonymous reviewers.

References

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

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*, 105–116.

Bruce, V. (1988). *Recognising faces*. Hove, UK: Erlbaum.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*, 339–360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*, 207–218.

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

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

Bruck, M., Cavanagh, P., & Ceci, S. J. (1991). Fortysomething: Recognizing faces at one's 25th reunion. *Memory and Cognition*, *19*, 221–228.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science*, *10*, 243–248.

Campbell, R., Coleman, M., Walker, J., Benson, P. J., Wallace, S., Michelotti, J., & Baron-Cohen, S. (1999). When does the inner-face advantage in familiar face recognition arise and why? *Visual Cognition*, *6*, 197–216.

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.

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., & Milne, A. (1982). Recognizing faces in and out of context. *Current Psychological Research*, *2*, 235–246.

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

Davies, G. M., Shepherd, J. W., & Ellis, H. (1978). Remembering faces: Acknowledging our limitations. *Journal of Forensic Science*, *18*, 19–24.

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

de Haan, E. H., & Hay, D. (1986). The matching of famous and unknown faces, given either the internal or external features: A study on patients with unilateral brain lesions. In H. D. Ellis, F. Jeeves, F. Newcombe, & A. W. Young (Eds.), *Aspects of face processing* (pp. 302–309). Dordrecht, The Netherlands: Martinus Nijhoff.

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

Frowd, C. D., Bruce, V., Ness, H., Thomson-Bogner, C., Peterson, J., Mcintyre, A., & Hancock, P. J. B. (in press). Parallel approaches to composite production. *Ergonomics*.

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

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

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

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

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

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

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*, 445–464.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 986–1004.

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

Lee, K. L., & Perrett, D. I. (1997). Presentation-time measures of the effect of manipulations in colour space on discrimination of famous faces. *Perception*, *26*, 733–752.

Memon, A., & Bruce, V. (1983). The effects of encoding strategy and context change on face recognition. *Human Learning*, *2*, 313–326.

O'Donnell, C., & Bruce, V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*, *30*, 755–764.

Rattner, A. (1988). Convicted but innocent: Wrongful conviction and the criminal justice system. *Law and Human Behavior*, *12*, 283–293.

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

Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips: Everyday difficulties in recognising people. *British Journal of Psychology*, *76*, 495–523.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*, 737–746.

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

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

Received 7 June 2005; revised version received 8 February 2006