A Decade of Evolving Composite Techniques: Regression- and Meta-Analysis

Charlie Frowd (1*), William B. Erickson (2), James M. Lampinen (2), Faye C. Skelton (3), Alex H. McIntyre (4) and Peter J.B. Hancock (4)
(1) Department of Psychology, University of Winchester, Winchester SO22 4NR
(2) Department of Psychological Science, University of Arkansas, Fayetteville AR 72704
(3) School of Psychology, University of Central Lancashire PR1 2HE UK
(4) Psychology, School of Natural Sciences, University of Stirling, Stirling FK9 4LA
* Corresponding author: Charlie Frowd, Department of Psychology, University of Winchester, Winchester SO22 4NR, UK. Email: Charlie.Frowd@winchester.ac.uk. Phone: (01962) 624943.

Abstract
Purpose – The article assesses the impact of seven variables that emerge from laboratory research involving facial-composite construction using popular police systems: EvoFIT, Feature and Sketch.
Design/methodology/approach – The paper involves regression- and meta-analyses on composite-naming data from 23 studies that have followed procedures used by police practitioners for forensic face construction. The corpus for analyses contains 6464 individual naming responses from 1,069 participants in 41 experimental conditions.
Findings – The analyses reveal that composites constructed from the holistic EvoFIT system were over four-times more identifiable than composites from Feature and Sketch systems; Sketch was somewhat more effective than Feature systems. EvoFIT was more effective when internal features were created before rather than after selecting hair and the other external features. The holistic cognitive interview (H-CI) was shown to promote a valuable improvement (cf. CI) in naming for the three system types tested. The analysis also confirmed that composites were considerably less effective when constructed from a long (1 - 2 day) compared with a short (0 - 3.5 hour) retention interval.
Originality/value – A range of variables were assessed that are of importance to forensic practitioners who construct composites with witnesses and victims of crime. The main results are that EvoFIT using an internal-features method of construction is superior, as is the H-CI administered prior to face construction for three contrasting production systems.
Keywords facial composite, EvoFIT, feature system, sketch, H-CI, regression, meta-analysis.
Paper type Research paper

Around 15 years ago, we were aware of a prevalent view among forensic practitioners: procedures used to construct composites had been largely optimised and the effectiveness of a composite was determined by the ability of the witness. The procedures used to construct composites in a forensic setting were clearly involved, and these are described in detail in Fodarella, Kuivaniemi-Smith and Frowd (2015, this volume). In brief, for a traditional 'feature' system, a practitioner would administer cognitive-interviewing (CI) techniques, to obtain a description of the offender's face from a witness (who may also be a victim), and then prepare an 'initial' composite: a face with facial features (eyes, nose, mouth, etc.) to match the description. The practitioner would then present alternative features from the software system for the witness to select the best-matching items, with selected features adjusted for most appropriate size and placement; finally, a paint package would be used to add lines, wrinkles, etc. The aim here is to allow a witness to achieve the best likeness possible of the offender. Alternatively, a forensic artist would produce a composite sketch, usually by hand, following a similar procedure. The artist would first obtain a description of the offender's face from the witness (via a CI) and prepare an 'initial', faintly-drawn sketch. Artist and witness would then work together on the configuration properties of the face (spacing of facial features), and subsequently to increase the level of detail of the features themselves. In either case, the resulting composites would be shown to other people (police officers and members of the public) to identify.

To quantify the effectiveness of composites used in this way, Frowd, Carson, Ness, McQuiston et al. (2005) defined a 'gold' standard by which composite systems (or new techniques) should be assessed in the laboratory: composite construction should follow procedures used in police interviews and performance of systems (or techniques) should be based on people's ability to spontaneously name the resulting composites. Using this procedure, what a decade of research has revealed is that fairly-good performance is possible when the interval is up to a few hours in duration from encoding a target face to constructing a composite of it. Here, constructors using sketch and modern feature systems prevalent in the US, UK and Europe (e.g. E-FIT, PRO-fit, FACES, Identikit 2000) create composites that other people name with a mean of around 20% correct (e.g., Brace, Pike and Kemp, 2000; Bruce et al., 2002; Frowd, Carson, Ness, McQuiston et al., 2005). However, when the retention interval is a day or two in duration, a usual minimum in police investigations, mean correct naming of the resulting composites is usually low (M = ~5%) (e.g., Frowd, Bruce, Ness et al., 2007; Frowd, Carson, Ness, Richardson et al., 2005; Frowd,

McQuiston-Surrett et al., 2007). Thus, the procedures being used for face construction seemed to be neither effective nor optimal.

Considerable research effort has sought solutions (to face construction) which attempt to be more-closely aligned to face recognition (a holistic process) than face recall (describing a face). As we tend to recognise faces as complete entities (wholes) rather than by their component parts (facial features) (e.g. Tanaka and Farah, 1993), face construction should be effective if implemented likewise. This concept has long been implemented in modern feature systems: individual features are presented for selection in the context of a complete face (Davies and Milne, 1982). For the emerging 'holistic' systems, this idea is taken a step further: constructors repeatedly select complete faces (or face regions) from arrays of alternatives, with characteristics of selected items being 'bred' together, to 'evolve' a composite. These systems also contain scales for witnesses to change age and other global properties of an evolved face. Overall, the approach is based on recognition, which is more stable over time than recall (Davies, 1983), and requires holistic processing of faces rather than explicit recall of facial features. Note the difference between recall and recognition: the process used with these systems does not preclude recognition of individual features but enhances the ability to utilise information more effectively (Tanaka and Farah, 1993). There are three main implementations: EvoFIT, which has been assessed extensively using the gold standard (e.g., Frowd, in press); EFIT-V (Gibson et al., 2009), evaluated using this standard in one published study (Valentine et al., 2010); and ID (Tredoux et al., 2006).

An important development that has led to a forensically-useful system (EvoFIT) turned out to relate to differences in which faces are processed when they are (i) constructed and (ii) named. The former is carried out by a witness or victim who is usually *unfamiliar* with a target (an offender), and in this case face processing is influenced strongly by perception of *external features* (hair, ears and neck): in contrast, internal features (the inner region encompassing eyes, brows, mouth, etc.) are particularly important for recognition of a familiar face (e.g. Bruce et al., 1999; Ellis, Shepherd and Davies, 1979; Fletcher, Butavicius and Lee, 2008; Young et al., 1985)—in this case, for successful naming of a composite by police officers and members of the public (Frowd, Bruce, Ness et al., 2007; Frowd et al., 2011). Frowd et al. (2010) demonstrated that using a Gaussian ('blur') filter to de-emphase the visual presence of external features in EvoFIT face arrays helped constructors to create composites with fairly-good correct naming (M = 25%) after a 2-day retention interval: composites with even-higher naming (M = 45%) emerged when external features were

masked completely until internal features had been constructed (Frowd, Skelton, Atherton, Pitchford, Hepton et al., 2012); for an example face array, see Fodarella, Frowd et al. (2015, this volume).

A further substantial development has been made by facilitating holistic processing prior to face construction, in this case by encouraging constructors to focus on the global appearance of a target (Frowd, Nelson et al., 2012): after constructors have freely recalled a target face, they reflect silently on the character of the face for one minute and then make seven whole-face judgements—such as its perceived level of honesty or masculinity. This so-called Holistic-CI[(i)] (H-CI) is also effective for feature-based composites (Frowd et al., 2008), and for artists' sketches (Kuivaniemi-Smith et al., unpublished) when facial features are selected in the context of a complete face (see Discussion). Furthermore, naming can be improved still further when finished composites from these systems are viewed (i) as a dynamic caricature (e.g. Frowd, Bruce, Ross et al., 2007), a process which exaggerates and then de-emphasises distinctive aspects of the face, and (ii) from side-on (e.g. Davis et al., 2015, this volume; Frowd, Jones et al., 2013).

It is worth mentioning at this stage that at least some of the aforementioned developments combine additively to increase naming. In Frowd, Skelton et al. (2013), EvoFIT composites constructed after a 24 hour retention interval, an H-CI and internal features first (masked external-features) were named side-on with a mean of 74% correct (and similar suspect identification has been found for EvoFIT in criminal cases: Frowd, Pitchford et al., 2012). This kind of performance is also possible from a feature system (see Discussion). Together, results indicate that it is now possible to construct highly-identifiable composites from contrasting systems. (For a more-detailed review of the literature, refer to Frowd, in press.)

To summarise, the approach of accessing memory by selection of face arrays with either blurred external features or internal features first produces a more-effective composite than by selection of individual facial features. The question is by how much, and how does this improve by masking external features? Similarly, what is the overall benefit of the H-CI? Answers to such questions should be of interest to forensic practitioners, to allow them to assess the effectiveness of composites created in criminal investigations, and for contributing to theories about how we construct and recognise faces.

Our main aim then is to quantify factors involved in face construction: interview (CI and H-CI), system (holistic, feature and sketch), EF (external-features blurring and masking) and associated factors (e.g. retention interval). Based on available and sufficient composite naming data from published and unpublished studies that have followed the gold-standard procedure, two main analyses are presented. First is a Logistic Regression on studies that have investigated system, interview and study characteristics. Second is a meta-analysis (to support the result from Logistic Regression) specifically looking at the advantage of the H-CI. We also provide possible direction for future research.

Method

The Composite Data Set

Research studies were considered for inclusion with designs that aimed to mimic the forensic use of composites. This necessitated that studies be published in the past decade, as this was when the gold standard was developed (Frowd, Carson, Ness, Richardson et al., 2005). To adhere to this standard, it was necessary that researchers involved in face construction: (a) did not see the target under construction, (b) were trained in cognitive-interviewing techniques and administered a CI (or H-CI) for participants to recall the appearance of a target face (see Coding section below on Interview), and (c) were trained on the relevant composite system and aimed to create the best possible likeness with participants in as much time as was necessary. At a minimum, researchers were trained 'in house' and then practiced extensively on interview and system before recruiting participants to construct composites. It was also important that our DV was spontaneous naming: while we acknowledge that other metrics have been used to evaluate the visual quality of composites (e.g. Bruce et al., 2002; Ellis, Shepherd and Davies, 1975; Frowd, Bruce, Ness et al., 2007), the ecological validity of composite systems can only be properly assessed via direct face recognition. Also pertinent to this standard were conditions that involved targets which were *unfamiliar* to constructors and were created after a minimum retention interval of one day. Projects with other study characteristics (SC) were considered, see following section, to allow preliminary analyses to be conducted on these variables.

It was important that there were at least four sets of naming responses per IV and SC, to allow computation of a stable estimate. This decision led to excluding data from EFIT-V, as there was only one appropriate set available (Valentine et al., 2000); FACES 3.0, as there were only two sets (Frowd, Carson, Ness, McQuiston et al., 2005; Frowd, McQuiston et al., 2007); and the archaic Photofit (Frowd, Carson, Ness, Richardson et al., 2005). Data were also excluded from early (non-commercial) prototypes of EvoFIT, in particular prior to

development of external-features blurring around 2006, since these experimental versions make it difficult to define a specific system. For further details on conditions of inclusion, see following section on Coding.

Composite naming data from 23 studies met these criteria for research emanating from the Universities of Stirling, Central Lancashire, Dundee and Winchester. Studies are summarised in Table 1. As can be seen, 15 studies collected data on EvoFIT, 15 on the PRO-fit and E-FIT feature systems, and four on Sketch; of these, two studies included a comparison between Feature and Sketch, and one between Feature and EvoFIT. Seven studies contributed data to more than one condition, and so are listed in separate rows in the table [e.g. FS13(a) and FS13(b)], while four (not necessarily mutually-exclusive) studies contributed to both CI and H-CI (FS13, FN12, FB08 and KSUP). Overall, there are 41 individual conditions.

The corpus contained full-data sets arising from publications in academic journals and proceedings of conferences; these included seven unpublished studies (N = 9 conditions), to limit overestimation of effect sizes (e.g. McLeod and Weisz, 2004). Twenty-seven composite researchers collected data for face construction and naming. Participants were adult (17+ years), fluent English speakers (since the interview involved face recall and this was carried out in English throughout). In more detail, there were 432 participants who constructed a single composite from memory with the assistance of a trained researcher. Once the composites had been constructed, of which there were between eight and 16 (M = 10.3, SD = 1.3) per study, these images were given to further participants to name. Naming was carried out by a total of 637 people, to provide a corpus of 6464 individual responses.

Coding

The primary DV was accurate naming, the most forensically-relevant measure. A numeric value of 1 was assigned when participants either correctly named a composite or provided an unambiguous semantic description of the face: in contrast, a value of 0 was assigned for either a wrong name or no name. For all included studies, after participants had been presented with their assigned set of composites, participants were asked to name the actual targets, to check that they were actually familiar with the relevant identities. In a small percentage of cases (M = 4.2%), the presented target was not correctly named, and so the associated composite was screened out (treated as missing data) since these items could not have been correctly named (i.e. the identity of such composites was unknown to the

participant). Note that this method of coding can produce a measure of central tendency which is different but very similar to mean values reported in the relevant papers[ii].

The second DV was inaccurate (mistaken) responses, since it is sensible to examine participants' willingness to offer names to composites in general. For this second DV, higher scores *per se* suggest less-accurate composites, since they may appear to be visually similar to one or more other identities. Mistaken names can, however, be of benefit to law enforcement by allowing potential suspects to be eliminated from an investigation. In signal-detection terminology, when correct names do not change, an increase in mistaken names can be considered as an increase in response *bias*, a representation that elicits more-frequent responding. Here, mistaken names were coded as 1, and 0 if no name was offered. Cases were again removed for unknown targets, as were composites that were correctly recognised (to give N = 3372 responses). As a measure of central tendency, the fraction incorrect i is defined as:

$$i = \frac{\text{number of wrong names}}{\text{number of wrong names + number of no names}} \qquad (1)$$

For example, in the first data row of Table 1, a name (either correct or incorrect) was given for almost all of the composites where the target was known: 67.2% of these were correct and, of the remaining 32.8%, 92.9% involved a wrong name and 7.1% were cases where no name was offered.

The data set emerged with sufficient responses (from N >= 4 study conditions) to consider three important independent variables (IVs) and four study characteristics (SCs):

1. *System* (IV). Four prevalent production systems were able to be included: EvoFIT, E-FIT, PRO-fit and Sketch. As E-FIT and PRO-fit are very similar in operation and effectiveness (e.g. see the two Frowd, Carson et al., 2005 papers), we refer to them generically as Feature systems. Similarly, sketches were created by three artists and were coded equivalently. System thus had three levels (1 = EvoFIT, 2 = Feature and 3 = Sketch), as illustrated in Figure 1; based on the aforementioned research (e.g. Frowd et al., 2010; Frowd, Skelton, Atherton, Pitchford, Hepton et al., 2012), EvoFIT was expected to produce composites with the highest correct naming.

Figure 1. Composites constructed in the included studies from (left to right) EvoFIT, Feature and Sketch systems. Images were produced by different constructors (in different studies) 24 hours after each person had seen a photograph of UK footballer, Frank Lampard. For copyright reasons, we are unable to reproduce the photograph itself; instead, an accurate likeness has been created, far right (courtesy of forensic artist, Heidi Kuivaniemi-Smith).

2. Interview (IV). The type of interview administered by the researcher prior to face construction was coded at two levels (1 = CI and 2 = H-CI)—see far-right pair of columns in the table. The CI is a flexible set of techniques (e.g. Wells, Memon and Penrod, 2007) and here involved a rapport-building stage, and further mnemonics for participants to: (i) think back to the time when their target had been seen and visualise the face (reinstatement of context), and (ii) recall as much information as possible about the face without guessing. Researchers did not interfere with this *free* recall exercise, except to ask participants to slow down if they spoke too fast for the researcher to write down the given information. The CI varied somewhat across studies, sometimes involving a second cycle of free recall (e.g. Frowd, Carson, Ness, McQuiston et al., 2005) and at other times when participants were invited to provide further information on their initial recall as part of *cued* recall (e.g. Frowd, Carson, Ness, Richardson et al., 2005). Such variation was not expected to substantially change identification of composite across conditions (see Frowd, Nelson et al., 2012 for a discussion on this issue). As described above, the H-CI involved character attribution after target face recall. Composites were expected to be superior following an H-CI than a CI.

3. *External Features, EF* (IV). Constructors created a composite with Feature and Sketch systems with external features visible. For (non-prototype) versions of EvoFIT, however, constructors repeatedly selected from arrays of faces presented to them in one of two ways. The first method involved arrays presented with blurred external-features (Blur). The second, which research indicates is a more effective method (e.g. Frowd, Skelton, Atherton, Pitchford, Hepton et al., 2012), involves arrays revealing internal features only

(IF) (with external features selected at the end of the construction process). EF type (1 = EF Blur and 2 = IF) was assessed in a separate analysis for EvoFIT composites.

4. Target Mode (SC). Targets were presented to constructors in colour as a photograph or video (1 = photograph and 2 = video). The latter mode involved a target (i) speaking into the camera or (ii) engaged in an interaction with another person in a naturalistic setting (e.g. a café); constructors listened to video clips with headphones. Two previous meta-analyses (Meissner and Brigham, 2001; Shapiro and Penrod, 1986) reported no reliable effect of mode of presentation on recognition hits, and so we predicted the same null outcome for correct naming of composites. Clearly, use of target videos is more forensically-valid than photographs (and we consider this issue in more detail in the Discussion).

5. Target Source (SC) varied considerably (see Table 1, Source). A preliminary analysis of the average naming rates suggested that composites were less identifiable for identities generally in the public eye (labelled 'Celebrity' in Table 1, $N = 7$) than non-celebrity targets, and so target source was coded dichotomously (1 = non celebrity and 2 = celebrity).

6. Retention interval (SC) spanned 0 (immediate construction), 3-to-4 hours, 20-to-28 hours and 44-to-52 hours. Correct naming of composites was expected to decline with (face construction following) increasing delay, but not as a linear function (e.g. a greater decline from 0 to 1 day than from 1 to 2 days, as would be predicted by Ellis, Shepherd and Davies, 1980; cf. Ebbinghaus, 1885), and so coding was *short* (0 hours, $N = 4$), *medium* (3 - 4 hours, $N = 6$) and *long* (20 to 52 hours, $N = 31$). Obviously, the *long* group is most forensically relevant.

7. Foil composites (SC). The final variable was related to simulating conditions of real-life use of composites rather than to forensic practice. Fourteen conditions contained from two to 10 'foil' composites in the testing set presented to participants for naming. These foils were of *unfamiliar* identities, not from the target set. Participants were warned of their presence from the start, the aim being to limit identification by a process of elimination and to encourage recognition of the actual composites; the presence of foils also reflects the real-world situation where a composite (e.g. when seen in the newspapers) will not always be a face that is familiar to an observer. Logically, foils should inhibit adoption of a lax response criterion, or elevated response bias. This aim is supported by Shapiro and Penrod (1986) who report fewer misidentifications (false alarms) in the presence of foils (decoys). As for retention interval, presence of foils was quantised (1 = absent and 2 = present).

Exclusions. While it would have been potentially valuable to consider a predictor for duration of target encoding (see Discussion), few conditions varied from 60 seconds for photographs, and so this SC was not included in the analysis. For the same reason,

offenders are sometimes known to witnesses, and a composite of them can still be important when there is uncertainty regarding identity—such as in cases of deception. While some research indicates appreciable benefit to face construction of prior familiarity with a target (e.g. Davies et al., 2000; Frowd et al., 2011), there were insufficient cases available to allow inclusion of data for which targets were familiar to constructors. So, all studies involved construction of an unfamiliar target[iii]. This issue of insufficient data was handled in the same way (not included) for conditions involving (i) unconventional face databases (using sketch-like vs. photographic features), (ii) unconventional presentation of target stimuli (in greyscale vs. colour), (iii) constructors who were invited to make a decision in an unusual way (rapid selection of faces), (iv) unconventional construction (sequential presentation of faces in EvoFIT arrays), (v) constructors who were subjected to a stress intervention at encoding, and (vi) targets who were not white Caucasian.

Table 1. Characteristics of studies included in the analyses.

Study	*System*	*EF*	*Target*			*Naming*				
			Mode	Source	Delay (hr)	Foils	CI		H-CI	
FNUP	EvoFIT	Blur	Photo	Football	0	0	67.2	(92.9)		
FS13(a)	EvoFIT	Blur	Photo	TV Soap	24	0	24.1	(42.6)	42.5	(50.0)
FN12(a)	EvoFIT	Blur	Photo	Football	24	4	17.6	(22.5)	32.5	(36.4)
FL09	EvoFIT	Blur	Photo	Football	24	0	21.3	(5.3)		
FS12(a)	EvoFIT	Blur	Photo	Football	24	4	23.4	(44.6)		
FN12(b)	EvoFIT	Blur	Video	Retail	24	5	24.1	(65.1)	39.6	(69.6)
FN12(c)	EvoFIT	Blur	Video	Retail	24	4	22.5	(68.8)	35.8	(46.8)
HB11	EvoFIT	Blur	Photo	Uni/staff	24-48	10	26.5			
FOUP	EvoFIT	Blur	Photo	Retail	48	2	24.4	(66.2)		
FP10(a)	EvoFIT	Blur	Photo	Snooker	48	0	21.6	(27.6)		
FF15	EvoFIT	IF	Photo	TV Soap	24	0	41.8	(22.8)		
FEUP	EvoFIT	IF	Photo	TV Soap	24	4	37.8	(60.9)		
FS13(b)	EvoFIT	IF	Video	TV Soap	24	0	36.7	(46.0)	53.8	(43.2)
FS12(b)	EvoFIT	IF	Photo	Football	24	4	45.9	(40.3)		
FDUP	EvoFIT	IF	Photo	Football	24	0	44.8	(35.5)		
FTUP(a)	Feature	Vis.	Photo	Football	0	0	27.5	(39.7)		
FS11	Feature	Vis.	Photo	Football	0	0	31.9			
FB07	Feature	Vis.	Photo	Uni/staff	0	8	17.5			
FR05(a)	Feature	Vis.	Photo	Celebrity	3.5	0	22.4			
FR05(b)	Feature	Vis.	Photo	Celebrity	3.5	0	16.0			
FB08	Feature	Vis.	Video	TV Soap	3.5	0	8.6	(69.4)	41.2	(65.4)
FTUP(b)	Feature	Vis.	Photo	Football	3.5	0	7.5	(35.1)		
FM05(a)	Feature	Vis.	Photo	Celebrity	48	0	0.0			
FM05(b)	Feature	Vis.	Photo	Celebrity	48	0	1.5			
FM07	Feature	Vis.	Photo	Celebrity	48	0	1.1	(7.5)		
FN07	Feature	Vis.	Photo	Football	48	0	4.2	(50.7)		
FTUP(c)	Feature	Vis.	Photo	Football	48	0	11.3	(39.4)		
FF11	Feature	Vis.	Photo	Football	48	0	1.3	(9.3)		
PS06	Feature	Vis.	Photo	Football	48	0	3.1	(35.0)		
FP10(b)	Feature	Vis.	Photo	Snooker	48	0	4.1	(35.5)		
FR05(c)	Sketch	Vis.	Photo	Celebrity	3.5	0	9.8			
SAUP	Sketch	Vis.	Video	TV Soap	24	4	15.8	(68.8)		
KSUP	Sketch	Vis.	Photo	Football	24	4	14.3	(45.2)	23.5	(40.4)
FM05(c)	Sketch	Vis.	Photo	Celebrity	48	0	6.9			
EMMeans†										
1	EvoFIT 56.0	Blur 29.0	Photo 29.0[a]	Non celeb. 33.0	0 hr 42.0[a]	No foils 30.0	CI 19.0		H-CI 37.4	
2	Feature 14.7	IF 50.1	Video 25.8[a]	Celebrity 23.4	3.5 hr 35.6[a]	Foils 26.2				
3	Sketch 21.7				1-2 day 12.3					

Note. Figures are in percentage for accurate naming and, where available, for inaccurate naming in

parentheses; see text for their calculation. For conciseness, a succinct code for each Study has been created: see list of References for definitions. For *EF* (external features), the coding was whether this region was visible (Vis.), blurred (Blur) or masked (IF, internal features only) at face construction. For Source, targets for (a) *Football* were UK international-level footballers, (b) *Retail* were staff working in retail outlets, (c) *Uni/staff* were staff working at a university, (d) *Snooker* were professional snooker players and (e) *Celebrity* were famous faces.

†Estimated Marginal Means (EMMeans) are percentage-correct naming by numerically-coded category. All contrasts for predictors are significant, $p < .02$, except for column-wise [a] $.05 < p < .10$. See Endnote [iv] for calculation of EMMeans (listed at the bottom of the table) given the associated Odds Ratio.

Logistic Regression

The principal analyses involved Logistic Regression mainly due to superior statistical power relative to other approaches (e.g. ANOVA). Separate analyses were carried out (using SPSS version 21) on accurate- and inaccurate-naming responses: for Model A with all variables entered except for EF, and for Model B, to assess EF for EvoFIT composites.

Validity checks: For both of these models, the usual checks were conducted for goodness-of-fit-based tests: $f > 0$, and $f(expected) < 5$ for no more than 20% of cells. It is worth noting that Model B (EvoFIT) has fewer observations and lower statistical power. In fact, during checks of validity (e.g. Field, 2009), and partly due to the reduction in model size, target source was of concern. No issues were apparent for Model A (Collinearity: predictors' *VIF* < 1.6 and *Tolerance* > .7, eigenvalues were sensible in the scaled cross-products matrix; dependencies were not strong between variables; and residual errors were independent, Durbin-Watson $1.5 < DW < 2.0$). For Model B, however, retention interval and target source were strongly related, and collinearity was an issue for source (*VIF* = 8.5, *Tolerance* = .1), and so source was not included. Also, data from the short-delay condition (FNUP) were excluded as there were insufficient responses to sensibly examine retention interval.

Once models were built, standard errors (SE) of Beta (B) coefficients were checked for improbable (too low or too high) values. Also, the fit of points were verified as appropriate (less than 2.4% of cases had Studentized residuals > 2, and < 0.1% were > 2.5), and no points exerted undue influence (*Cook's Distance* < 0.03; *Leverage* ~=3*(k+1)/n; $0.01 < |DFBeta|(\max) < 0.14$), indicating model stability.

Model A. Full model (includes all variables except for EF type)

Accurate naming. The analysis commenced with a saturated model, one containing all predictors except for EF, with IVs and SCs subject to backward-sequential removal ($p > .1$) based on the Likelihood Ratio[v]. On this occasion, the six variables (i.e. all except for EF) were reliable predictors of accurate naming and were included in the model (Table 2). For each predictor, the lowest numerically-coded category was used as reference, and B coefficients reflect this scheme. For instance, interview was referenced to CI (coded as 1) and, as H-CI (2) promoted more identifiable faces, *B* is positive: in contrast, *B* is negative for source, as non-celebrity targets (1) produced more identifiable faces than celebrity targets (2). System was a trichotomous IV, and two contrasts relative to EvoFIT (the lowest category) indicated superiority relative to (i) Feature and (ii) Sketch; a third contrast (iii) revealed benefit of Sketch over Feature. There was a reliable deficit in naming for retention interval (trichotomous IV) from short to long, and from medium to long; the deficit from short and medium approached significance.

Any reliable increase in correct naming of composites would be welcomed in criminal investigations, but a worthwhile gain occurs when *Exp(B)* > 2—that is, for predictors which at least double naming rates. An *Exp(B)* of around 2 is interpretable as a 'medium' effect size by Sporer and Martschuk (2014), but we argue (as do Morris and Fritz, 2013) that effect sizes should be domain specific: for facial composites, this gain should be considered as 'large' as it is a useful effect for policing, *Exp(B)* of 1.5 as 'medium' and 1.2 as 'small'. Using these guidelines, large effects occurred for EvoFIT (cf. Feature and Sketch), the H-CI (cf. CI) and for long (cf. short and medium) delays. Mode, Source and Foils exerted much-weaker effects. To aid interpretation, Estimated Marginal Means (EMMeans) are presented for each variable at the bottom of Table 1 (see also *Note* for this table).

Table 2: Accurate naming for the full Logistic-Regression model.

Variable	*N*	*B*	*SE(B)*	X^2	*DF*	*p*	*Exp(|B|)*
System				315.88	2	< .001	
i. EvoFIT > Feature	16	-2.01	0.11	312.21	1	< .001	7.4 [6.0, 9.3]
ii. EvoFIT > Sketch	5	-1.54	0.17	84.66	1	< .001	4.6 [3.3, 6.5]
iii. Sketch > Feature	5	-0.47	0.16	8.97	1	.003	1.6 [1.2, 2.2]
Interview: H-CI > CI	7	0.94	0.09	103.98	1	< .001	2.5 [2.1, 3.1]
Mode: Photograph > Video	11	-0.16	0.09	2.97	1	.09	1.2 [1.0, 1.4]
Source: Non-Celebrity > Celebrity	7	-0.48	0.15	10.71	1	.001	1.6 [1.2, 2.2]
Retention interval				192.42	2	< .001	
i. Short > Medium	4	-0.27	0.15	3.36	1	.07	1.3 [1.0, 1.8]
ii. Short > Long	4	-1.64	0.13	157.19	1	< .001	5.2 [4.0, 6.7]

iii. Medium > Long	6	1.37	0.14	103.76	1	< .001	3.9	[3.0, 5.1]
Foil composites: None > Foils	15	-0.19	0.08	5.48	1	.019	1.2	[1.0, 1.4]
Constant		-0.96	0.08	148.49	1	< .001	2.6	

Note. Model [$X^2(8) = 724.3$, $p < .001$, Cox and Snell $R^2 = .11$, Nagelkerke $R^2 = .17$]. The following is presented for each variable: Beta (B) coefficient (slope of the variable's regression line), standard error of *B* (*SE(B)*), Wald Chi-square (X^2), degrees of freedom (*DF*), model fit (*p*), Odds Ratio (*Exp(|B|)*) and [in square brackets] ±95% CI for *Exp(|B|)*. For ease of interpretation, the Odds Ratio (effect size) is always expressed as a value greater than 1.0, by taking the exponential of the absolute value of *B*, rather than allowing it to appear as a Risk Ratio (a value less than 1.0). For probability values, the (APA) convention is followed by expressing non-significant contrasts ($p > .05$) to 2 d.p. *N* is the minimum number of comparisons involved in the calculation; interview, for instance, has *N* = 7 as there are 34 conditions for CI and 7 for H-CI.

Inaccurate naming. For this DV, higher inaccurate (mistaken) names *per se* indicate less-accurate composites. The analysis followed the same basic procedure as above. System was removed in Step 1 ($p = .28$), and the final model is summarised in Table 3. For interview, while accurate naming greatly increased, the H-CI led to fewer inaccurate names—the ideal forensic case. Each categorical increase in retention interval (i.e. from short to medium, and from medium to long) roughly halved the rate of inaccurate names given, and is somewhat similar to the reduction in accurate naming—essentially, a decrease in response bias. While photographs led to composites with slightly more frequent correct names than videos, inaccurate names were much-less frequent, indicating superiority for use of photos. Celebrity (vs. non-celebrity) stimuli reduced accurate and (to a much-greater extent) inaccurate names—again, a decrease in response bias. While it was expected that foil composites would avoid a liberal response criterion, the opposite effect seemed to be occurring: naming involving foil composites markedly *increased* inaccurate responses.

Table 3: Inaccurate naming for the full Logistic-Regression model.

Variable	*N*	*B*	*SE(B)*	X^2	*DF*	*p*	*Exp(|B|)*	
Interview: CI > H-CI	7	-0.24	0.10	5.66	1	.017	1.3	[1.0, 1.5]
Mode: Video > Photograph	11	0.94	0.10	95.13	1	< .001	2.6	[2.1, 3.1]
Source: Non-Celebrity > Celebrity	7	-1.49	0.40	14.08	1	< .001	4.5	[2.0, 9.7]
Retention interval				92.11	2	< .001		
i. Short > Medium	4	-0.65	0.24	7.51	1	.01	1.9	[1.2, 3.0]
ii. Short > Long	4	-1.51	0.22	48.89	1	< .001	4.5	[2.9, 6.8]
iii. Medium > Long	5	0.86	0.12	52.35	1	< .001	2.4	[1.9, 3.0]
Foil composites: Foils > None	15	0.84	0.09	85.17	1	< .001	2.3	[1.9, 2.8]
Constant		-0.21	0.22	0.89	1	.35	1.2	

Note. Model [$X^2(7) = 462.4$, $p < .001$, Cox and Snell $R^2 = .13$, Nagelkerke $R^2 = .17$]. For definition of variables, see Table 2, *Note*.

Model B. Model for EvoFIT (includes all variables except for system)

Accurate naming. There were 2539 accurate responses to EvoFITs, of which 5.4% were screened out (again for targets which were not correctly named, but also for responses from the short-delay condition, as explained above). Source was removed in Step 1 ($p = .24$) and foils in Step 2 ($p = .32$); Table 4 summarizes the final model. There was a sizeable benefit for IF (cf. blur) construction, and the H-CI benefit was similar to that found in Model A.

Table 4: Accurate naming for Logistic Regression model for EvoFIT composites.

Variable	*N*	*B*	*SE(B)*	X^2	*DF*	*p*	*Exp(|B|)*	
External Features (EF): IF > Blur	7	0.90	0.10	86.42	1	< .001	2.5	[2.0, 3.0]
Interview: H-CI > CI	5	0.61	0.10	33.85	1	< .001	1.8	[1.5, 2.2]
Constant		-0.44	0.06	61.20	1	< .001	1.5	

Note. Model [$X^2(2) = 104.0$, $p < .001$, Cox and Snell $R^2 = .04$, Nagelkerke $R^2 = .06$]. See Table 2, *Note*.

Inaccurate naming. The model for inaccurate EvoFIT naming is summarised in Table 5. IF (cf. blur) construction led to composites with somewhat higher inaccurate responses; the other variables produced effects that were consistent with those found in Model A.

Table 5: Inaccurate naming for Logistic Regression model for EvoFIT composites.

Variable	*N*	*B*	*SE(B)*	X^2	*DF*	*p*	*Exp(|B|)*	
External Features (EF): IF > Blur	5	0.48	0.14	11.62	1	.001	1.6	[1.2, 2.1]
Interview: CI > H-CI	6	-0.24	0.14	2.73	1	.10	1.3	[1.0, 1.7]
Mode: Video > Photograph	6	1.03	0.12	69.11	1	< .001	2.8	[2.2, 3.6]
Foil composites: Foils > None	7	1.21	0.13	92.98	1	< .001	3.3	[2.6, 4.3]
Constant		-0.25	0.08	10.65	1	.001		

Note. Model [$X^2(4) = 180.1$, $p < .001$, Cox and Snell $R^2 = .11$, Nagelkerke $R^2 = .15$]. See Table 2, *Note*.

Meta-Analyses of Interviewing Styles

In addition to system, interviewing style has attracted considerable attention in recent composite research: in particular, comparing CI with the recently-developed H-CI. Since interview has been examined across several ($N = 7$) conditions, we were able to conduct a

more-traditional meta-analysis to estimate the magnitude of the overall effect, the results of which were expected to be similar to and thus support those of the Logistic Regression.

There is a dearth of meta-analyses for composite data. The most relevant to the current work is Meissner and Brigham (2001) who revealed that the process of constructing a composite increased people's ability to identify a target (by 1.6 times): here, we assess the extent to which the holistic component of the interview improves other people's ability to identify a composite.

Meta-analyses have a unit of analysis at the level of the individual study (rather than at the level of the participant or item). They estimate the existence and magnitude of effects while accounting for "noise" within different studies, in particular for the random-effects model (used here) which assumes inter-study variability. These analyses also take into account the fact that larger samples tend to provide more-accurate estimates of the corresponding populations—that is, the sampling error of the effect size tends to be reduced for larger than for smaller samples. We have followed procedures of Lipsey and Wilson (2001) and, as SPSS does not have the inherent functionality, conducted analyses using a Microsoft Excel template made available by Neyeloff, Fuchs and Moreira (2012).

Method

Studies. As before, the same seven comparisons (from $N = 4$ studies) comparing CI and H-CI were used; DVs were participant responses to composites for which the relevant target had been correctly named.

Procedure. Names given to composites are dichotomous (correct or incorrect) and so effect sizes for meta-analyses are appropriately expressed as the weighted logged Odds Ratio, OR_{logged}. In the above regression analysis, these two response types were compared with no-name responses, to give accurate and inaccurate measures. In the meta-analysis, the same important comparison was conducted for accurate naming, but we also compared accurate with inaccurate, to provide an estimate of the overall naming advantage of the H-CI.

Effect sizes were first conducted by calculating the odds ratios (*OR*s) for each interviewing outcome according to Equation (2):

$$OR = \frac{p(treatment\ 1\ Response\ A)/\ p(treatment\ 1\ Response\ B)}{p(treatment\ 2\ Response\ A)/\ p(treatment\ 2\ Response\ B)} \qquad (2)$$

The remaining calculations require values to be centred on zero. However, *OR*s are centred on one, and so the natural log of the *OR*s was calculated, to give OR_{logged}. The resulting values were then aggregated, assuming a random-effects model, as each study examined different factors in addition to interview style: unrelated idiosyncratic differences among studies (see Table 1) result in normally-distributed differences in effect sizes that are due to more than simple sampling error within studies. For each comparison below, we present Fisher's *Z* and associated *p*-value for the measured effect; OR_{logged} and its ±95% confidence interval; and the *OR* itself, which is $Exp(OR_{logged})$, and is analogous to *Exp(B)* used in the above regression analyses.

Results

Accurate naming. The main analysis contained 1489 correct-name and no-name responses, and detailed results are shown in the Forest plot in Figure 2. See Neyeloff et al. (2012) for interpreting this type of graph—briefly, a square indicates the odds ratio for a particular study with an area that is proportional to size of the effect, and horizontal lines that indicate 95% confidence intervals. Interview was reliable [$Z = 3.20$, $p < .001$, $OR_{logged} = -0.82$, *CI95*(-1.32, -0.32)], with an effect size [*OR* = 2.3] that is very similar to the one measured in Model A [*Exp(B)* = 2.5], supporting the overall superiority of H-CI over CI by correct naming.

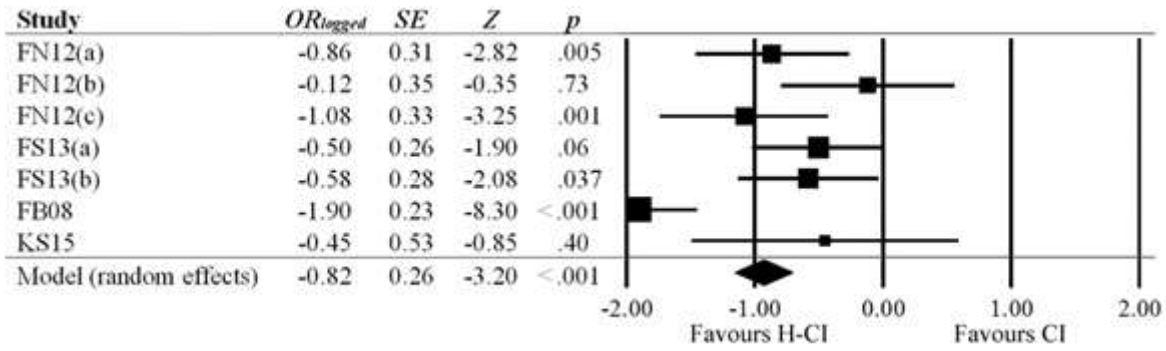

Figure 2. Forest plot of the H-CI versus CI advantage (OR_{logged}) for accurate naming.

Accurate versus inaccurate naming. This analysis contained 1468 responses that were either correct or incorrect. Interview was again reliable [$Z = 3.02$, $p < .001$, $OR_{logged} = -0.89$, *CI95*(-1.47, -0.31), *OR* = 2.4], indicating substantial overall benefit for accurate over

inaccurate naming.

Discussion

It is crucial that law enforcement obtain effective composites from witnesses and victims, to allow offenders to be apprehended promptly. An important exercise, then, is to assess the effectiveness and reliability of techniques for constructing composites. Here, we assembled a corpus of identification data from 23 composite studies with procedures that were aligned to forensic face construction and naming. Logistic Regression analyses confirmed the large advantage of (i) the H-CI (cf. CI), supported by the meta-analysis, and (ii) EvoFIT, both overall and when using the internal-features (cf. EF blur) method of construction.

It is evident that accessing memory by the EvoFIT approach is effective: the benefit to correct naming was over four times that of either feature or sketch. One advantage of EvoFIT is that composites can be created readily even when witnesses cannot recall an offender's face—although if they can, an H-CI can be administered, boosting performance (Frowd, Skelton et al., 2013). Facial information is forgotten rapidly (e.g. Ellis et al., 1980), and arguably contributes to the decline in utility of feature systems with increasing delay. Here, increases in retention interval led to a less-accurate representation (Model A), faces with much-lower accurate and inaccurate naming. This indicates an overall reduction in response bias, suggesting that composite images become more generic (less like any specific identity) over longer retention intervals—which is not surprising as this variable also affects recognition (e.g. Shapiro and Penrod, 1986). While other feature systems should likewise produce ineffective composites after long delays, as indicated by Frowd, McQuiston et al. (2007), there is currently insufficient data to be confident of the rate of decline with sketch. Ongoing research is charting the change in naming over time by system, including delays upward of a week, which are somewhat common in forensic practice (e.g. Frowd, Pitchford et al., 2012).

The work also confirms the benefit of EvoFIT composites created using the IF method: while incorrect names increased somewhat using this procedure relative to EF blur, correct names increased to a greater extent. When first applied to a feature system, this novel IF method did not generalise: in fact, it did the opposite, reducing correct naming of composites. More-recent work, however, reveals a large benefit in naming for IF construction when the interview uses an H-CI rather than a face-recall CI (manuscript in preparation). Taken together, it seems that a side-effect of the H-CI is to shift a

constructor's attention from the whole face to the internal features, allowing IF construction to become effective after an H-CI. This suggestion does makes good sense since global judgements and impressions incorporated in the H-CI are less likely to be influenced by external features of a target face. Similarly, the H-CI was not initially beneficial for practitioners who sketch composites (Stops, unpublished). As mentioned at the start of the paper, traditional sketch production involves witnesses describing features of the face (using a CI) and an artist drawing the face to create an 'initial' sketch; witnesses then request changes be made to this face. What appears to be important for the H-CI to be effective (cf. CI) is that constructors must select features in the context of a complete face (which is how feature systems usually operate, Skelton et al., this volume) rather than carrying out what is essentially a recall-based task: to comment upon the likeness of an initial sketch. Indeed, sketches created in this way (by context-based selection of facial features) following an H-CI were included in the current analyses (KSUP).

The accurate-naming data (Model A) also indicated that sketch is somewhat-more effective [*Exp(B)* = 1.6] than a feature system, the advantage of which has been observed previously (e.g. Frowd, Carson, Ness, McQuiston et al., 2005; Laughery and Fowler, 1980). Sketch production does seem to involve an important qualitative difference in task performance: constructors tend to work on groups of features rather than sequentially on individual features, and so forensic sketching should be aligned somewhat closer to holistic face processing (Davies and Little, 1990; Laughery, Duval and Wogalter, 1986). Our work also highlights that there is minimal naming data available on the effectiveness of sketches, and future research could address this issue along with quantifying individual differences between artists, which are known to exist using other dependent measures of composite quality (e.g. Laughery and Fowler, 1980). We were also able to resolve one issue raised by Frowd, Carson, Ness, Richardson et al. (2005): some sketches have limited detail, potentially causing confusion about the intended identity. There was no evidence of this concern, however, as inaccurate naming did not vary reliably by system. In this case, all three types of system created composites that were wrongly named to the same extent.

Unfortunately, there were insufficient naming data (correct or incorrect) to be able to sensibly consider the other holistic system in forensic use, EFIT-V (Gibson et al., 2009), software that uses a somewhat similar face selection and breeding process to EvoFIT. EFIT-V has not been evaluated extensively by naming, but one study, Valentine et al. (2010), reveals that spontaneous naming of individual composites was 20.3% correct

(targets were videos of TV soap actors, a CI was administered, retention interval was *short*, and no foils were used). This mean is similar to naming of feature composites constructed likewise (*M* = 26.6% correct for FTUP(a), FS11 and FB07) after a short retention interval. As EvoFITs are named considerably higher than this estimate even after a long retention interval (e.g. EvoFIT IF construction in Table 1), it is unlikely that EFIT-V is as effective. This may, in part, be due to EFIT-V presenting face arrays with intact external features: neither external-features blurring nor IF construction are used, both of which have been shown to increase performance (e.g. here; Frowd, Skelton et al., 2013). Future research could establish whether this is indeed the case, how this system performs under forensically-relevant conditions (esp. after a long retention interval) and whether the H-CI is effective.

The remaining variables concerned study characteristics. Accurate naming marginally favoured targets presented as photographs than videos: there was a null effect for EvoFIT, presumably due to reduced power for Model B. This result suggests that use of static photographs in laboratory research is a sensible proxy to the more-realistic use of video stimuli when the DV of interest is correct naming. However, mistaken names for videos were much higher both overall and for EvoFITs, indicating encoding superiority of photos. While videos are clearly closer to real life, presentation of photos for short encoding duration does parallel a real-life application, such as when an offender's face is seen briefly from a specific viewpoint. The photos used depicted to target in a largely front-face view, the same view as the included composite systems, and so should provide a condition for optimal face construction (Frowd et al., 2014)—although this may not be the best angle of view in itself for unfamiliar identities (see Ness et al., this volume). The situation is similar to more-accurate recognition of unfamiliar faces when the angle of view is the same (cf. different) between study and test (e.g. Bruce, 1982). With videos, it is perhaps the fine detail in a target face that is not as well reproduced in a composite, leading to a representation that is more-easily confused with another identity (hence the large increase with inaccurate names). Certainly the benefit of feature versus more-global (holistic) encoding is established for face construction (e.g. Frowd, Bruce, Ness et al., 2007; Wells and Hryciw, 1984). It should also be the case that encoding duration (not able to be assessed here due to insufficient data) should be positively related to correct naming (as it is to face recognition, e.g. Shapiro and Penrod, 1986) and, to a greater extent in the opposite direction, to incorrect naming; future work could explore the impact of this forensically-useful variable. Further, in particular for use of video stimuli, there is evidence that audio

can be distracting to a constructor, in particular if it involves incongruent speech (e.g. Marsh, Skelton and Frowd, 2015, this volume).

Similarly, people in the public eye (e.g. well-known celebrities) are sometimes employed as targets, and our work reveals that their involvement yields composites with somewhat lower correct naming but much higher mistaken naming. One explanation is that we are aware of more celebrities than identities from any other category, and so simply have more names to offer: we are likely to be familiar with hundreds of celebrities, but far fewer top UK football players. Future work might therefore explore the relationship between potential size of target pool and frequency of name production. Recent research (as yet unpublished), however, hints that an alternative account may be related to attractiveness, a facial property which is generally higher for celebrity than non-celebrity targets. The new research reveals that lower-attractiveness targets emerge as more identifiable (even when controlling for relevant factors such as distinctiveness), a result which fits with the current finding. Ongoing research is attempting to resolve which of these explanations is likely to be correct.

In relation to the first of these accounts, researchers usually exercise caution if the target pool is limited, such as when stimuli are staff from a university department: a warning is given to (naming) participants that not all composites are of a specified category (e.g. department staff) and foil composites are introduced into the testing set. The aim is to avoid naming by a process of elimination and to encourage recognition. We have confirmed that the procedure with foils suppresses correct naming of composites (Model A), although the effect size was small: inaccurate naming was much higher with foils than without, a result that runs counter to their influence in face-recognition studies (Shapiro and Penrod, 1986). It may simply be that participants naming composites become less discriminative in general after they know that foils are present, prompting them to offer more names and be less-accurate overall. It is currently unknown, however, whether these effects are being driven by prior warning of foils, or their actual presence. With this question in mind, we re-ran the regression for Model A but included foils as a continuous rather than a discrete variable: foils remained a reliable predictor for both DVs, but the effect size was noticeably reduced for inaccurate naming [*Exp(B):* 2.3 to 1.2]. This suggests that the number of foils is important, and so both mechanisms may be at play. Future research could inform on this methodological issue.

In summary, the project sought a better understanding of the effectiveness of facial composites. The approach involved a corpus collected over the past decade using composite naming as the main measure of assessment. There were over six thousand individual naming responses in over 40 experimental conditions, and the analyses revealed some interesting results. It is clear that the holistic EvoFIT system creates composites with over four times higher correct naming than those from the feature and sketch systems tested; it was also found that the EvoFIT approach is much-more effective when external features are masked rather than blurred in face arrays. Use of the holistic component to interviewing and a shorter (cf. longer) retention interval also promoted more-identifiable composites. Milder benefits to naming emerged for use of video stimuli (cf. photos) and without use of so-called 'foil' composites. Ongoing work is exploring the impact of retention interval for various systems, the impact of facial attractiveness, and target-pool size at naming.

References

(Codes in square brackets are studies included in logistic-regression and/or meta-analyses.)

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

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

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*, Vol. 87, pp. 894-902.

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences*, 2nd ed. New York, Academic Press.

Davies, G.M. (1983), "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 Little, M. (1990), "Drawing on memory: Exploring the expertise of a police artist", *Medical Science and the Law*, Vol. 30, pp. 345-354.

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, pp. 119-124.

Ellis, H.D., Shepherd, J.W. and Davies, G.M. (1975), "An investigation of the se of the photo-fit technique for recalling faces", *British Journal of Psychology*, 66, pp. 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*, Vol. 8, pp. 101-106.

Davis, J. et al. (2015), “An Evaluation of post-production facial composite enhancement techniques”, *Journal of Forensic Practice.*

Field, A. (2009), *Discovering statistics using SPSS*, 3rd Ed., Sage, London.

Fletcher, K.I., Butavicius, M.A. and Lee, M.D. (2008), Attention to internal face features in unfamiliar face matching, *British Journal of Psychology*, Vol. 99, pp. 379–394.

[FF15]Fodarella, C., Hepton, G., Stone, K., Date, L. and Frowd, C.D. (2015), “Split-face construction using the holistic EvoFIT system”, *Journal of Forensic Practice*.

Fodarella, C., Kuivaniemi-Smith, H.J. and Frowd, C.D. (2015). “Detailed procedures for forensic face construction”. *Journal of Forensic Practice*.

Frowd, C.D. (in press), “Facial composites and techniques to improve image recognisability”, in T. Valentine and J. Davis (Eds.) *Forensic facial identification: theory and practice of identification from eyewitnesses, composites and cctv*. Wiley-Blackwell.

[FB07]Frowd, C.D., Bruce, V., McIntyre, A. and Hancock, P.J.B. (2007), “The relative importance of external and internal features of facial composites”, *British Journal of Psychology*, Vol. 98, pp. 61-77.

[FN07]Frowd, C.D., Bruce, V., Ness, H., Bowie, L., Thomson-Bogner, C., Paterson, J., McIntyre, A. and Hancock, P.J.B. (2007), “Parallel approaches to composite production”, *Ergonomics*, Vol. 50, pp. 562-585.

Frowd, C.D., Bruce, V., Ross, D., McIntyre, A. and Hancock, P.J.B. (2007), “An application of caricature: how to improve the recognition of facial composites, *Visual Cognition*, Vol. 15, pp. 1-31.

[FB08]Frowd, C.D., Bruce, V., Smith, A. and Hancock, P.J.B. (2008), “Improving the quality of facial composites using a holistic cognitive interview”, *Journal of Experimental Psychology: Applied*, Vol. 14, pp. 276 – 287.

[FM05]Frowd, C.D., Carson, D., Ness, H., McQuiston, D., Richardson, J., Baldwin, H. and Hancock, P.J.B. (2005), “Contemporary Composite Techniques: the impact of a forensically-relevant target delay”, *Legal and Criminological Psychology*, Vol. 10, pp. 63-81.

[FR05]Frowd, C.D., Carson, D., Ness, H., Richardson, J., Morrison, L., McLanaghan, S. and Hancock, P.J.B. (2005), “A forensically valid comparison of facial composite systems”, *Psychology, Crime and Law*, Vol. 11, pp. 33-52.

[FDUP]Frowd, C.D. and Duckworth, L. (unpublished), “The impact of composite hair changes”.

[FEUP]Frowd, C.D., Erickson, W.B., Lampinen, J.M., Marsh, J.E., Coultas, C., Kneller, W. and Brown, C. (unpublished), “The impact of weapons and unusual objects on recall and composite construction”.

[FF11]Frowd, C.D. and Fields, S. (2011), “Verbalisation effects in facial composite production”, *Psychology, Crime and Law*, Vol. 17, pp. 731-744.

Frowd, C.D., Jones, S., Forarella, C., Skelton, F.C., Fields, S., Williams, A., Marsh, J., Thorley, R., Nelson, L., Greenwood, L., Date, L., Kearley, K., McIntyre, A. and Hancock, P.J.B. (2013), “Configural and featural information in facial-composite images”, *Science and Justice*, DOI: 10.1016/j.scijus.2013.11.001.

[FL09]Frowd, C.D., Lee, C., Petkovic, A., Nawaz, K. and Bashir, Y. (2009), “Further Automating and Refining the Construction and Recognition of Facial Composite Images”, International *Journal of Bio-Science and Bio-Technology*, Vol. 1, Vol. 59-74.

[FM07]Frowd, C.D., McQuiston-Surrett, D., Anandaciva, S., Ireland, C.E. and Hancock, P.J.B. (2007), "An evaluation of US systems for facial composite production", *Ergonomics*, Vol. 50, pp. 1987–1998.

[FNUP]Frowd, C.D., Miller, N. et al. (unpublished), "Morphing of EvoFIT composites".

[FN12]Frowd, C.D., Nelson, L., Skelton F.C., Noyce, R., Atkins, R., Heard, P., Morgan, D., Fields, S., Henry, J., McIntyre, A. and Hancock, P.J.B. (2012), "Interviewing techniques for Darwinian facial composite systems", *Applied Cognitive Psychology*, Vol. 26, pp. 576-584.

[FP10]Frowd, C.D., Pitchford, M., Bruce, V., Jackson, S., Hepton, G., Greenall, M., McIntyre, A. and Hancock, P.J.B. (2010), "The psychology of face construction: giving evolution a helping hand", *Applied Cognitive Psychology*, Vol. 25, pp. 195-203.

Frowd, C.D., Pitchford, M., Skelton, F.C., Petkovic, A., Prosser, C. and Coates, B. (2012), "Catching Even More Offenders with EvoFIT Facial Composites", In A. Stoica, D. Zarzhitsky, G. Howells, C. Frowd, K. McDonald-Maier, A. Erdogan, and T. Arslan (Eds.) *IEEE Proceedings of 2012 Third International Conference on Emerging Security Technologies* (pp. 20 - 26). DOI 10.1109/EST.2012.26.

Frowd, C.D., Skelton, F.C., Atherton, C., Pitchford, M., Bruce, V., Atkins, R., Gannon, C., Ross, D., Young, F., Nelson, L., Hepton, G., McIntyre, A.H. and Hancock, P.J.B. (2012), "Understanding the multi-frame caricature advantage for recognising facial composites", *Visual Cognition*, 20, pp. 1215-1241.

[FS12]Frowd, C.D., Skelton F., Atherton, C., Pitchford, M., Hepton, G., Holden, L., McIntyre, A. and Hancock, P.J.B. (2012), "Recovering faces from memory: the distracting influence of external facial features", *Journal of Experimental Psychology: Applied*, Vol. 18, pp. 224-238.

[FS11]Frowd, C.D., Skelton, F., Butt, N., Hassan, A. and Fields, S. (2011), "Familiarity effects in the construction of facial-composite images using modern software systems", *Ergonomics*, Vol. 54, pp. 1147-1158.

[FS13]Frowd, C.D., Skelton F., Hepton, G., Holden, L., Minahil, S., Pitchford, M., McIntyre, A., Brown, C. and Hancock, P.J.B. (2013), "Whole-face procedures for recovering facial images from memory", *Science and Justice*, Vol. 53, pp. 89-97.

[FOUP]Frowd, C.D., Thompson, R. (unpublished), "Combining holistic and feature based composite construction".

[FTUP]Frowd, C.D. and Tran, L. (unpublished), "Feature based face construction over increasing retention interval".

Frowd, C.D., White, D., Kemp, R.I., Jenkins, R., Nawaz, K. and Herold, K. (2014), "Constructing faces from memory: the impact of image likeness and prototypical representations", *Journal of Forensic Practice*, Vol. 16, pp. 243-256.

Gibson, S.J., Solomon, C.J., Maylin, M.I.S. and Clark, C. (2009), "New methodology in facial composite construction: from theory to practice", *International Journal of Electronic Security and Digital Forensics*, Vol. 2, pp. 156-168.

[HB11]Hancock, P.J.B., Burke, K. and Frowd, C.D. (2011), "Testing facial composite construction under witness stress", *International Journal of Bio-Science and Bio-Technology*, Vol. 3, pp. 65-71.

[KSUP]Kuivaniemi-Smith and Frowd, C.D., "Improving the effectiveness of sketch-based composite images".

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

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

Lipsey, M.W. and Wilson, D. (2001), *Practical Meta-Analysis*, Sage, London.

Marsh, J., Skelton, F.C. and Frowd, C.D. (2015), Distractibility: congruent versus incongruent speech at face encoding, *Journal of Forensic Practice*.

McLeod, B.D. and Weisz, J.R. (2004), "Using dissertations to examine potential bias in child and adolescent clinical trials", *Journal of Consulting and Clinical Psychology*, Vol. 72, pp. 235–251.

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

Morris, P.E. and Fritz, C.O. (2013), "Effect sizes in memory research", *Memory*, doi:10.1080/09658211.2013.763984.

Ness, H. et al. (2005), "Are two views better than one? A study examining recognition of three-quarter view and full-face composites", *Journal of Forensic Practice*.

Neyeloff, J.L., Fuchs, S.C. and Moreira, L.B. (2012), "Meta-analyses and Forest plots using a microsoft excel spreadsheet: step-by-step guide focusing on descriptive data analysis", *BioMed Central Research Notes*, doi: 10.1186/1756-0500-5-52.

[PS06]Plews, S. (2006), "The influence of some factors affecting Facial composite production and their Application in practical policing", MPhil dissertation, University of Stirling.

[SAUP]Stops, A. (unpublished), "Production techniques for sketching", MSc Forensic Art dissertation, University of Dundee.

S.L. Sporer, N. and Martschuk. (2014), "The Reliability of Eyewitness Identifications by the Elderly: An Evidence-based Review", In (Eds.) Michael P. Toglia, David F. Ross, Joanna Pozzulo, Emily Pica. *The Elderly Eyewitness in Court*, Psychology Press, New York.

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

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

Wells, G.L. and Hryciw, B. (1984), "Memory for faces: encoding and retrieval operations", *Memory and Cognition*, Vol. 12, pp. 338-344.

Wells, G.L., Memon, A. and Penrod, S.D. (2007), "Eyewitness evidence: improving its probative value", *Psychological sciences in the public interest*, 7, pp. 45-75.

[i] To be precise, this enhanced interview involves the standard techniques (mnemonics) of the cognitive interview (Frowd, in press), for witnesses to recall the appearance of the face (e.g. rapport building, reinstatement of context and free recall), and two additional mnemonics: silent recall of the target's personality and whole-face character attribution. As the literature (e.g. Frowd et al., 2008) refers to the overall technique as the holistic cognitive interview, or H-CI, we have followed this convention.

[ii] A frequently-used measure is the Conditional Naming Rate. CNR is calculated as the number of composites correctly named divided by the relevant number of targets correctly named; it can be computed by-participants and by-items, and is usually subjected to ANOVA. For examples, see Frowd, Carson, Ness, Richardson et al. (2005) and Valentine et al. (2010). When differences by target familiarity are minimal, the uncorrected (unconditional) naming rate is sometimes reported (e.g. Brace et al., 2000; Frowd et al., 2008). Alternatively, for the same reasons as ours, recent research (e.g. Frowd, Atherton, Pitchford, Hepton et al., 2012) has tended to analyse naming responses with regression techniques.

[iii] All included studies contained a pre-screening phase as a check that targets were *unfamiliar* to constructors. This involved participants having a quick look at a (randomly-selected) target; if the identity was familiar, another face was presented likewise, and participants encoded the first face which was unfamiliar to them. In contrast, a post-screening phase checked that targets were *familiar* to participants engaged in naming. This time, participants were presented with target images after having seen the relevant set of composites (for naming). Also, studies applied an *a priori rule*: participants needed to name most of the targets (typically $M >$ 75%) for their data to be included, or another participant would be recruited as a replacement.

[iv] If n is percentage-correct naming for one condition, the fraction correct $p = n / 100$, and the odds that a composite will be correctly named $P' = [p / (1 – p)]$. Similarly, if m is percentage-correct naming in an associated condition, the fraction correct $q = m / 100$, and the odds $Q' = [q / (1 – q)]$. The Odds Ratio $OR = P' / Q'$ or $[p / (1 – p)] / [q / (1 – q)]$. Rearranging, $m = P' / [OR + P'] * 100$. For example, from Table 1, Column 2, for EvoFIT, $n = 56.0$, $P' = [.56 / (1 - .56)] = 1.273$, OR (EvoFIT to Feature) = 7.4, and so naming m(Feature) = 1.273 / [7.4 + 1.273] * 100 = 14.7%.

[v] As a check of consistency, the regression models were run without backward elimination. The initial (saturated) models contained the same reliable predictors with virtually identical coefficients and effect sizes.