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

Charlie D. Frowd (1*), William B. Erickson (2), James M. Lampinen (2), Faye C. Skelton (3), Alex H. McIntyre (3) 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 72701
(3) Psychology, School of Life, Sport & Social Sciences, Edinburgh Napier University EH14 1DJ
(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. Email: Charlie.Frowd@winchester.ac.uk. Phone: (01962) 624943.

Frowd, C.D., Erickson, W.B., Lampinen, J.L., Skelton, F.C., McIntyre, A.H., & Hancock, P.J.B. (in press). A decade of evolving composite techniques: regression- and meta-analysis. *Journal of Forensic Practice.*

Abstract

Purpose – The article assesses the impact of seven variables that emerge from forensic research on facial-composite construction and naming using contemporary 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' (E-FIT and PRO-fit) 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 (blurred) external features. Adding questions about the global appearance of the face (as part of the Holistic-Cognitive Interview, H-CI) gives a valuable improvement in naming over the standard face-recall Cognitive Interview (CI) for all 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 – Variables were assessed that are of importance to forensic practitioners who construct composites with witnesses and victims of crime. The main result is that EvoFIT using the internal-features method of construction is superior; an H-CI administered prior to face construction is also advantageous (cf. face-recall CI) for EvoFIT as well as for two further contrasting production systems.

Fifteen years ago, a prevalent view among forensic practitioners was that 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 detailed (described in Fodarella *et al.*, 2016), with the aim of allowing a witness (who may also be a victim) to create the best likeness of an offender. In brief, for traditional 'feature' systems, a practitioner would administer cognitive-interviewing (CI) techniques, to obtain a description of the offender's face from a witness, and then prepare an 'initial' composite: a face with facial features (eyes, nose, mouth, etc.) to match this description. Next, the practitioner would present alternative features from the software system for the witness to select best-matching items, with selected features adjusted for size and placement. Finally, a paint package could be used to add lines, wrinkles, etc. Alternatively, a forensic artist would produce a composite sketch. The artist would obtain a description of the offender's face from a witness (via a CI) and prepare an 'initial', faintly-drawn sketch. Artist and witness would work together on the configural properties of the face (spacing of features), and then to increase the overall level of detail. 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, Frowd *et al.* (2005a) 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 composite effectiveness should be based on people's ability to spontaneously name these images. Using this procedure, a decade of research has revealed that fairly good performance emerges when the interval is up to a few hours in duration from encoding a target face to constructing a composite of it. 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 *et al.*, 2000; Bruce *et al.*, 2002; Frowd *et al.*, 2005a). However, when the retention interval is one or two days, a usual minimum in police investigations, mean correct naming is usually low ($M \approx 5\%$; e.g., Frowd *et al.*, 2005b, 2007d). Thus, procedures used for face construction seemed to be neither effective nor optimal.

Considerable effort has sought solutions which are more closely aligned to face recognition (a holistic process) than to face recall (describing a face). As we tend to recognise faces as complete entities rather than by component parts (facial features) (e.g., Davies and Milne,

1982; Tanaka and Farah, 1993), face construction should be effective if accomplished likewise. This concept has long been implemented in modern feature systems: individual features are presented for selection in the context of a complete face (Skelton *et al*., 2016). For the emerging 'holistic' systems, this concept is taken one step further: constructors repeatedly select whole faces (or whole-face regions) from arrays of alternatives, with characteristics of selected items being 'bred' together, to 'evolve' a composite. They also contain scales for changing 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 features. There are three main implementations: EvoFIT, which has been assessed extensively using the gold standard (Frowd, 2015); EFIT-V (Gibson *et al.*, 2009), evaluated using the gold standard in one published study (Valentine *et al.*, 2010); and ID (Tredoux *et al.*, 2006).

A crucial observation that led to a forensically-useful system, EvoFIT, concerns differences regarding the way in which faces are processed when constructed and named. Face construction is performed by a witness who is usually *unfamiliar* with a target (an offender), and so a witness's processing of the face is influenced strongly by *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 *et al.*, 1979; Young *et al.*, 1985)—in this case, for successful naming of a composite (Frowd *et al*., 2007a, 2011). Frowd *et al.* (2010) used a Gaussian ('blur') filter to de-emphasise external features in EvoFIT arrays. They demonstrated that this technique helped constructors to create composites with fairly good correct naming (M = 25%) after a two-day retention interval, presumably as this prevented external features from dominating during construction of an unfamiliar face. Composites with even higher naming (M = 45%) were produced when just internal features were shown, with external features added thereafter (Frowd *et al.*, 2012d); for an example face array, see Fodarella *et al.* (2016).

A further important development was made by facilitating holistic processing prior to face construction: after witnesses have freely recalled a target face using CI techniques (e.g., Wells *et al.*, 2007), they reflect silently on its character for one minute and then make seven whole-face judgements—such as its level of perceived honesty or masculinity. These two whole-face techniques, when used after a face-recall CI, form the Holistic-Cognitive interview (H-CI). Constructors then build the face as normal. The H-CI improves

composite naming from EvoFIT (Frowd *et al.*, 2012a), feature systems (Frowd *et al.*, 2008) and artists' sketches (Kuivaniemi-Smith and Frowd, unpublished, see Discussion).

Naming is improved still further when composites are viewed (i) as a dynamic caricature (e.g., Frowd *et al.*, 2007c), an image format that exaggerates and de-emphasises distinctive aspects of the face, and (ii) from side-on, to allow the face to appear long-and-thin (e.g., Davis *et al.*, 2016; Frowd *et al.*, 2013a). Some of the aforementioned developments are also complimentary, and combine to increase naming substantially. In Frowd *et al.* (2013b), EvoFIT composites, constructed after a 24 hour retention interval using the H-CI and masked external features, were named side-on with a mean of 74% correct (and a similar level of identification has been found for EvoFIT in criminal cases: Frowd *et al.*, 2012b). Such performance is also possible from feature systems (see Discussion). Together, these results indicate that it is now possible to construct highly-identifiable composites from contrasting systems.

To summarise, the approach of accessing memory by selection from face arrays with blurred external features produces more effective composites than by selection of individual facial features. The question is by how much, and how does this improve when external features are masked in the face arrays during construction? Similarly, what is the overall benefit of the H-CI? Answers to questions such as these 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 (independent variables, IVs) 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 involving studies that have investigated system, interview and study characteristics. Second is a meta-analysis looking at interview. Direction is provided 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 were conducted in the past 10 years, as this was when the gold standard was developed (Frowd *et al.*, 2005b). To adhere to the standard, it was necessary that researchers involved in face construction: (a) did not see the target under construction, so as not to inadvertently influence the participant, (b) were trained in cognitive-interviewing techniques and administered CI (or H-CI) for participants to recall the appearance of a target face, and (c) were trained on the relevant composite system and aimed to create the best likeness possible with participants without time constraints. At a minimum, researchers were trained 'in house' and practiced extensively on interview and system prior to constructing composites for the relevant study. It was also important that the primary measure (the dependent variable, DV) was spontaneous naming: while other metrics have been used to assess the visual quality of composites (e.g., Bruce *et al.*, 2002; Ellis *et al.*, 1975; Frowd *et al.*, 2007b), the ecological validity of composite systems can only be properly assessed via direct face recognition. Also pertinent to this standard were constructors who were *unfamiliar* with the target identities and created a composite 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.

We also required at least four sets of naming responses for each IV or SC, to allow computation of stable estimates. This requirement led to exclusion of EFIT-V, as only one set was available (Valentine *et al.*, 2000); FACES 3.0, as there were only two sets (Frowd *et al.*, 2005a, 2007d); and the archaic Photofit (Frowd *et al.*, 2005b). Data were also excluded from non-commercial prototypes of EvoFIT, specifically prior to development of external-features blurring around 2006, since these experimental versions make it difficult to define a specific system. See following section for further details of criteria for inclusion.

Composite naming data from 23 studies met these main criteria for research emanating from the Universities of Stirling, Central Lancashire, Dundee and Winchester. As can be seen in Table 1, 15 studies involved data on EvoFIT, 15 on 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 studies contributed to both CI and H-CI (FS13, FN12, FB08 and KSUP). There

are 41 individual conditions, summarised in the table as 34 rows for CI and another seven for H-CI (far right column).

The corpus comprised full-data sets described in academic journals and proceedings of conferences, and, to limit overestimation of effect sizes (e.g., McLeod and Weisz, 2004), from seven unpublished studies (N = 9 conditions). Twenty-seven trained researchers administered standard face-construction and face-naming procedures on 1,069 adult (17+ years) fluent-English-speaking participants. A total of 432 participants constructed a single composite from memory with the assistance of one of these researchers. Each study produced between eight and 16 composites (M = 10.3, SD = 1.3) per experimental condition. These composites were then presented sequentially to a further 637 participants to name. The set contained 6464 individual naming responses.

Coding and exclusions

The primary DV was accurate naming. A value of 1 was assigned when participants gave a correct name or an appropriate unambiguous semantic description for a composite: a value of 0 was assigned for an incorrect name or when a name was not given. For all included studies, after attempting to name their randomly-assigned set of composites, participants were invited to name a photograph of the targets, to establish familiarity with the relevant identities. When such a target was not correctly named, it was assumed that the participant would have been unable to accurately name the associated composite. In these cases (M = 4.2% overall), the relevant items were treated as missing data and not subject to analyses. Note that this coding scheme gives an estimate of central tendency that can be different but very similar to mean values reported in the relevant papers[i].

The second DV was inaccurate responses. Overall, an increase in the number of mistaken names *per se* indicates less accurate composites, images which tend to be similar to another identity. In signal detection terms, when correct and mistaken names increase at the same rate, this indicates an increase in response *bias*, a representation that elicits more frequent responding. From a forensic perspective, a mistaken name can generate a false lead; however, mistaken names are arguably less harmful than no names at all, since mistaken names provide a mechanism for potential suspects to be eliminated from an investigation.

Responses to composites were coded as 1 for wrong name, and 0 if no name was offered. Cases were again screened for incorrectly-named targets, but also for composite responses

that were named correctly (to give $N = 3372$ responses). As a measure of central tendency, the fraction *incorrect* is the number of wrong names divided by sum of wrong names and no names. In the first data row of Table 1, for instance, a name (correct or incorrect) was given for almost all composites when the target was familiar: 67.2% of these cases were correct and, of the remaining 32.8%, 92.9% were wrong names and 7.1% were no-name responses.

The available data set contained sufficient responses (for $N \geq 4$ individual conditions) to include three important independent variables (IVs) and four study characteristics (SCs):

1. *System* (IV). Four prevalent face-production systems were included: EvoFIT, E-FIT, PRO-fit and Sketch. E-FIT and PRO-fit are very similar in function (e.g., the Frowd *et al.*, 2005 papers), and are considered '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, 2012d), EvoFIT was expected to produce composites with highest correct naming.

Figure 1

2. *Interview* (IV). The CI included rapport-building, and mnemonics for participants to: (i) think back to the time of target encoding and visualise the face (reinstatement of context), and (ii) recall as much detail about the face as possible, without guessing. Researchers did not interfere with this *free*-recall exercise, except to ask participants to slow down if they spoke too fast for written notes to be made. The CI varied across studies, sometimes involving a second cycle of free recall (e.g., Frowd *et al.*, 2005a), or inviting elaboration (*cued* recall) on an initial account (e.g., Frowd *et al.*, 2005b). Such variation was not expected to noticeably change composites' identification (see Frowd *et al.*, 2012a for a discussion on this issue). The H-CI involved face-recall CI followed by character attribution. Type of interview (Table 1, far-right columns) was coded as 1 for CI and 2 for H-CI. Composites were expected to be superior following H-CI than CI.

3. *External Features, EF* (IV). Constructors traditionally create a composite using Feature and Sketch systems with external features always present. For EvoFIT, they repeatedly select from arrays of faces presented in one of two ways. In the first, external features appear blurred (Blur); in the second, which research suggests is more effective (e.g., Frowd *et al.*, 2012d), arrays contain internal features only (IF) and external features

are chosen towards the end of construction. EF type (1 = EF Blur and 2 = IF) was thus 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 person (i) speaking into the camera or (ii) interacting with another person in a natural setting (e.g., café); participants listened to video clips on headphones. Two meta-analyses (Meissner and Brigham, 2001; Shapiro and Penrod, 1986) report no reliable effect of mode of presentation on recognition hits, and so the same null outcome was predicted for correct naming of composites. Clearly, presentation is more forensically valid for videos than photographs.

5. *Target Source* (SC) varied considerably. A preliminary analysis of the data suggested that composites were less effective for well-known identities in the public eye ('Celebrity' in Table 1, $N = 7$), an effect which might be due to larger target-pool size (see Discussion), 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 increasing delay between target encoding and face construction, but not as a linear function (e.g., a greater decline from 0 to 1 day than from 1 to 2 days, based on Ellis *et al.*, 1980; cf. Ebbinghaus, 1885). Coding was *short* (0 hours, $N = 4$), *medium* (3 - 4 hours, $N = 6$) and *long* (20 - 52 hours, $N = 31$). The *long* interval is most forensically relevant in current practice.

7. *Foil Composites* (SC). The final variable concerned laboratory naming of composites. Fourteen conditions included from two to 10 'foil' composites; foils were of *unfamiliar* identities, not from the target set. Participants were warned of their presence, the aim being to avoid naming by a process of elimination and reflect real-world use: composites are not always of a familiar identity. Foil use should inhibit a lax response criterion, a prediction supported by Shapiro and Penrod (1986), who report fewer misidentifications (false alarms) for presence of foils (decoys). Foil use was dichotomised (1 = absent and 2 = present).

Exclusions. While duration of target encoding is interesting to study, few conditions varied from 60 seconds for photographs, and so this SC was not included. Similarly, offenders are sometimes a familiar identity (to a witness), and a composite can be useful in cases of uncertain identity (e.g., for confidence crimes). While research indicates sizeable benefit for construction of familiar targets (e.g., Davies *et al.*, 2000; Frowd *et al.*, 2011), data were again insufficient to allow analysis by target familiarity; indeed, all studies constructed an unfamiliar target[ii]. Likewise, not included were conditions with (i) unconventional face-databases (sketch-like features), (ii) unconventional presentation mode

of target stimuli (greyscale), (iii) constructors asked to make unusual decisions (rapid face selection), (iv) unconventional construction (sequential presentation of arrays), (v) constructors subjected to a stress intervention at encoding, and (vi) non-white targets.

Table 1

Logistic Regression

The principal analyses used Logistic Regression due to superior statistical power (cf. ANOVA). Separate analyses were conducted (using SPSS version 21) on accurate- and inaccurate-naming responses: for Model A involving all variables except EF, and for Model B, to assess EF for EvoFIT composites.

Validity checks. For both models, usual checks were made for a goodness-of-fit test: *f(observed)* > 0, and *f(expected)* < 5 for ≤ 20% of cells. No issues of validity (Field, 2009) were apparent for Model A (Collinearity: predictors' *VIF* < 1.6 and *Tol.* > .7, eigenvalues were sensible in the scaled cross-products matrix; dependencies were not strong between variables; and residual errors were independent, 1.5 < *Durbin-Watson* < 2.0). For Model B (EvoFIT), the variable Target Source was not included due to collinearity (*VIF* = 8.5, *Tol.* = .1), and Retention Interval was not included due to insufficient data; also, responses to composites from the single short-delay condition FNUP were excluded (due to their sizeable impact on accurate naming).

Models' Beta coefficients and their standard errors were checked for improbable (too low or high) values, and the fit of points was confirmed appropriate (< 2.4% of cases had Studentized residuals > 2, and < 0.1% were > 2.5); no points exerted undue influence (*Cook's Distance* < 0.03; *Leverage* ≈3*(k+1)/n; 0.01 < |*DFBeta*|(max) < 0.14), indicating stability.

Model A. All systems

Accurate naming. The analysis commenced with a saturated model containing all predictors except for EF, with IVs and SCs subject to backward-sequential removal ($p > .1$) based on Likelihood Ratio[iii]. All six predictors made a reliable contribution to accurate naming and so were included in the final model (Table 2). For each predictor, the lowest numerically-coded category was taken as reference (variables that are underlined in Tables 2 - 4), and Beta (B) coefficients reflect this scheme. CI (coded as 1) was reference for Interview and,

as H-CI (2) promoted more accurate faces, *B* is positive: *B* is negative for Source, as more identifiable faces emerged for non-celebrity (1) than celebrity targets (2). For System (trichotomous IV), contrasts indicated superiority of EvoFIT over (i) Feature and (ii) Sketch; a third contrast (iii) revealed benefit of Sketch over Feature. With Retention Interval (trichotomous IV), naming was higher for short than long, and for medium than long; the deficit from short to medium approached significance.

Any reliable increment in correct naming of composites would be welcomed in forensic practice, but a worthwhile benefit occurs when *Exp(B)* > 2—that is, for predictors which more than double naming rates. *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 composites, this gain should be considered 'large', due to impact for policing, with *Exp(B)* of 1.5 as 'medium' and 1.2 as 'small'. Based on these guidelines, large effects occur for EvoFIT (cf. Feature and Sketch), H-CI (cf. CI) and long (cf. short and medium) delays. For these three variables, the 95% confidence intervals of the effect size were narrow; also, the lower interval was large in size, indicating a substantial effect for the vast majority of likely true means. Mode, Source and Foils exerted much weaker effects. To aid interpretation, Estimated Marginal Means (EM_{Means}) are presented for each variable in Table 1 (see also *Note*).

Table 2

Inaccurate naming. Higher mistaken names *per se* indicate less accurate composites. The analysis followed the procedure as above. System was not a reliable predictor for this DV (p = .28) and so was removed in Step 1, yielding the final model (Table 3). For Interview, while accurate naming greatly improved with the H-CI, as found above, inaccurate names decreased—the ideal forensic outcome. Each categorical increase in retention interval (from short to medium, and from medium to long) roughly halved inaccurate names, which is somewhat similar to the decrease in accurate naming—essentially, a reduction in response bias. While target photographs promoted slightly more accurate composites than videos, inaccurate names were much less frequent, revealing superiority for photographs. Celebrity (vs. non-celebrity) stimuli reduced accurate and, to a much greater extent, inaccurate names. Lastly, foil composites were expected to inhibit a liberal response criterion, but the opposite emerged: foil use markedly *increased* inaccurate responses.

Table 3

Model B. EvoFIT

Accurate naming. There were 2539 accurate responses to EvoFITs, of which 5.4% were screened—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

Inaccurate naming. The EvoFIT model for inaccurate naming is also summarised in Table 4. IF (cf. blur) construction led to composites with somewhat higher inaccurate responses; the other variables produced effects consistent with those of Model A.

Meta-Analysis

There is a dearth of meta-analyses on facial composites. Arguably the most relevant is Meissner and Brigham (2001) who reveal that constructing a composite increases constructor's ability to identify a target (by 1.6 times). Here, we assess the extent to which holistic components of the interview improve the identifiability of a composite. Results were expected to be similar to and support those from the above Logistic Regression.

The unit of analysis for meta-analysis is at the level of the individual study rather than at the level of the participant, item or individual response. Meta-analyses 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 heterogeneity (inter-study variability). They assume that larger samples provide more accurate estimates of corresponding populations—that is, the error of the effect size tends to reduce for larger than for smaller samples.

We followed procedures of Lipsey and Wilson (2001) and (as SPSS does not have inherent functionality) conducted the meta-analyses using a modified version of the Microsoft Excel template made available by Neyeloff *et al.* (2012).

Studies. The same seven comparisons comparing CI and H-CI were used; DVs were participant responses to composites for which the relevant target had been correctly named.

Approach. Responses to composites are dichotomous (correct or incorrect) and so meta-analyses are expressed as the weighted logged Odds Ratio, OR_{logged}, an effect size analogous to *Exp(B)*. In our regression analysis, accurate responses were compared with no-name plus mistaken responses. This approach was followed for the meta-analysis, but we also directly compared accurate with inaccurate naming, to provide an estimate of the overall naming advantage of H-CI. Effect sizes were first obtained by calculating the odds ratios (*OR*s) for each interviewing outcome. The remaining calculations require values to be centred on zero and, as *OR*s are centred on one, the natural log of the *OR*s was used, to give OR_{logged}, and then aggregated, assuming a random-effects' model.

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 how to interpret this type of graph—briefly, a square indicates the odds ratio for a study with area proportional to size of the effect; horizontal lines indicate 95% CI. Interview was reliable [$Z = 3.20$, $p < .001$, $Q = 33.7$, $I^2(6) = 82.2\%$, $OR_{logged} = -0.82$], with an effect size (*OR* = 2.4, 95% CI [1.4, 4.0]) that is very similar to that measured in Model A (*Exp(B)* = 2.5 [2.1, 3.1]), supporting the superiority of H-CI over CI by correct naming. Note that the confidence intervals of the effect size are much narrower for the regression than the meta-analysis since the former is based on individual observations (rather than summary statistics), resulting in greater precision (for the regression analyses). Note also that between-study variability (heterogeneity I^2) is large, highlighting the presence of additional variability—as other factors are involved (e.g., system, retention interval).

Figure 2

Accurate versus inaccurate naming. This analysis contained 1468 responses that were correct versus mistaken (without no-name responses). Interview was reliable ($Z = 3.02$, $p < .001$, $OR_{logged} = -0.89$, $Q = 32.6$, $I^2(6) = 81.6\%$, *OR* = 2.4 [1.4, 4.4]), indicating a substantial overall advantage for H-CI over CI.

Discussion

It is crucial that law enforcement obtain effective composites from witnesses and victims, to

allow offenders to be apprehended promptly. Here, to assess the effectiveness of key stages in the process, a corpus of naming data was assembled from 23 studies using procedures that were aligned to forensic face construction and naming. The logistic-regression analysis confirmed a large advantage of (i) the H-CI (cf. CI), supported by the meta-analysis, and (ii) EvoFIT, both overall and using the internal-features (cf. EF blur) method of construction.

Accessing memory by EvoFIT is clearly effective: accurate naming was over four times that of Feature or Sketch (Table 2). Note that confidence intervals were fairly narrow, indicating a consistent, large estimate for this observation (even at the lower 95% CI); indeed, this same level of consistency occurred for all effect sizes for this DV, as one would expect using the current methodology of combining individual-response data from multiple studies. Another advantage of the EvoFIT approach (cf. traditional systems) is that witnesses are permitted to construct a composite even when they are unable to recall an offender's facial features (ACPO, 2009)—although if they can, an H-CI can be administered, facilitating performance (Frowd *et al.*, 2013b). Facial detail is forgotten rapidly (e.g., Ellis *et al.*, 1980), and this information loss arguably contributes to the decline in utility of feature systems with increasing delay. Here, longer retention intervals led to less accurate representations (Model A), faces with much lower accurate and inaccurate naming. This reduction in response bias suggests that faces are constructed more generically (less like any specific identity) with increasing delay— not surprising, as this variable also affects face recognition (e.g., Shapiro and Penrod, 1986). While other feature systems should likewise produce ineffective faces after long delays (e.g., Frowd *et al.*, 2007d), data are insufficient to be confident of the rate of decline for sketch. Ongoing research is charting naming rates by system, for delays upward of a week, which also occur in forensic practice (e.g., Frowd *et al.*, 2012b).

The work also confirms the benefit of the IF method of construction: while incorrect names increased using this procedure relative to EF blur, correct names increased to a greater extent. When first applied to a feature system, this IF method did not generalise: in fact, correct naming of composites reduced. More recent work, however, reveals a large benefit in naming for IF construction when using H-CI rather than CI (manuscript in preparation). It seems that a side-effect of H-CI is to shift a constructor's attention from the whole face to internal features, allowing IF construction to be effective after an H-CI. Similarly, H-CI was not initially effective for Sketch (Stops, unpublished). Using this method, witnesses usually describe facial features (via a CI) and a forensic artist draws an 'initial' sketch; they

then request changes to this face. What seems to be important for the H-CI (cf. CI) is that constructors select features in the context of a complete face (which is how feature systems usually operate, Skelton *et al.*, 2016) rather than carrying out what is essentially a recall task: to request changes to an initial sketch. Indeed, sketches created in this way (via whole-face feature selection) following an H-CI were included in our analyses (KSUP).

Results from Model A also established, in line with previous work (Laughery and Fowler, 1980), that Sketch is somewhat-more effective [*Exp(B)* = 1.6] than feature systems. Sketch production involves a potentially important qualitative advantage: witnesses tend to work on groups of features rather than on individual features, so allowing this forensic method to be closer aligned to holistic face processing (Davies and Little, 1990; Laughery *et al.*, 1986). Evidence was also provided to speak to an issue raised by Frowd *et al.* (2005b): some sketches have limited detail, potentially causing confusion about the intended identity. There was no evidence of this concern, as inaccurate naming did not vary reliably by system. In this case, all three types of system created composites that were mistakenly named to the same extent. The work did reveal that naming data were limited for Sketch, and research could address this issue along with quantifying individual differences between artists, which are known to exist (*ibid*.).

Unfortunately, even less naming data are available for the other holistic system in forensic use, EFIT-V (Gibson *et al.*, 2009), software that involves similar face selection and breeding to EvoFIT. EFIT-V has not been assessed extensively by naming, but one study, Valentine *et al.* (2010), reports naming of individual composites at 20.3% correct (targets were videos of TV soap actors, CI was administered, retention interval was short, and foils were not deployed). This mean value is comparable to naming of feature composites constructed likewise (*M* = 26.6% for FTUP(a), FS11 and FB07) after a short retention interval. As EvoFITs are correctly named at a much higher rate even after a long retention interval (EvoFIT IF construction, Table 1), EFIT-V is unlikely to be as effective. This may, in part, be due to EFIT-V showing face arrays with intact external features: neither EF blurring nor IF construction is used, both of which are effective (also confirmed here). Future research could establish whether this is indeed the case, how EFIT-V fares under forensically-relevant conditions (a long retention interval) and whether the H-CI is effective.

The remaining variables concern study characteristics. Accurate naming marginally favoured targets shown as photos rather than videos (the effect was null for EvoFIT,

presumably due to reduced power for Model B). This trend suggests that static images used for laboratory research are a good proxy to moving stimuli when the DV is correct naming. Mistaken names for target videos were much higher both overall and for EvoFITs, however, indicating encoding superiority for photos. While videos are closer to real life, short encoding of photos does parallel the situation where an offender's face is seen briefly. In addition, the photos in our constituent studies tended to present a frontal face, the same view as in the composite systems, and so stages of processing at construction overlap (Frowd *et al.*, 2014)—although perhaps not optimally for unfamiliar-face construction, as the best view may not be frontal (Ness *et al.*, 2016). For target videos, fine facial details may not be encoded as effectively as for photos, leading to composites that are more easily confused with other identities: hence the large increase in inaccurate names (Model A). Indeed, composites do seem to be more identifiable following feature (cf. more global) encoding (e.g., Frowd *et al.*, 2007b; Wells and Hryciw, 1984). It should also be the case that encoding duration (not assessed here due to insufficient data) is positively related to accurate naming, with the opposite effect for inaccurate naming, much as it is for face recognition (e.g., Shapiro and Penrod, 1986); by contrast, interference at encoding reduces composite quality (Marsh *et al.*, 2016). Future work could explore the impact of these forensically-relevant variables.

Targets in the public eye are sometimes used in lab studies at encoding, and our work reveals that using such well-known celebrities result in composites with lower correct and lower mistaken naming (although note that CIs for the latter DV were somewhat wider than elsewhere, indicating greater variability for the production of mistaken names). One explanation is that we are familiar with more celebrities than identities from any other category: we may be familiar with hundreds of celebrities, but far fewer top UK football players. For celebrities, this would create a higher density space of possible faces (cf. Lewis, 2004), leading to composites that are less effective as probes, suppressing name production. Future work might usefully explore the relationship between potential size of target pool and frequency of name production. Recent research (manuscript in preparation), however, hints that an alternative explanation may be related to attractiveness, a facial property which is normally higher for celebrity than non-celebrity targets. The research reveals that lower-attractiveness targets promote more identifiable composites (even when controlling for 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 explanation, researchers exercise caution if target-pool size is limited, such as when targets are staff from a university department: a warning is given to (naming) participants that not all composites are of a specific category (department staff) and foil composites are introduced into the testing set. The aim is to avoid naming by a process of elimination. We have confirmed that foil use suppresses correct naming (Model A), although the effect size was small. In contrast, inaccurate naming was much higher with foil use, a result that runs counter to their influence in face-recognition studies (Shapiro and Penrod, 1986). It may simply be that observers become less discriminative after they know that foils are present, prompting them to offer more names and be less accurate overall. It is currently unknown, however, whether this effect is being driven by prior warning of foils, or their actual presence. Future research could inform on this methodological issue.

To conclude, the project sought a greater understanding of the effectiveness of composites. A corpus of data was assembled from studies conducted over the last decade where naming was the dependent variable. The holistic EvoFIT system was found to produce composites with over four times higher correct naming than composites from Feature and Sketch systems; EvoFIT was also much more effective when external features were masked than blurred in face arrays; and composites were somewhat more identifiable from Sketch than Feature systems. Use of holistic components to cognitive interviewing and a shorter (cf. longer) retention interval both promoted more identifiable composites. Milder benefits to composite identification emerged for use of target photos (cf. videos) and without involving 'foil' composites. Ongoing work is exploring the impact of retention interval by system, 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.)

ACPO (2009), "Facial Identification Guidance", *National Policing Improvement Agency*.

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., 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.

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. and Milne, A. (1982), “Recognizing faces in and out of context”, *Current Psychological Research*, Vol. 2, pp. 235-246.

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.

Davis, J. Simmons, S., Sulley, L., Solomon, C. and Gibson, S. (2016), “An Evaluation of post-production facial composite enhancement techniques”, *Journal of Forensic Practice.*

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. (1979), “Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition”, *Perception*, Vol. *8*, pp. 431-439.

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.

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

[FF15]Fodarella, C., Hepton, G., Stone, K., Date, L. and Frowd, C.D. (unpublished), “Adjusting the focus of attention: Helping witnesses to evolve a more-identifiable composite”.

Fodarella, C., Kuivaniemi-Smith, H.J., Gawrylowicz, J. and Frowd, C.D. (2016). “Detailed procedures for forensic face construction”. *Journal of Forensic Practice*.

Frowd, C.D. (2015), “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, pp. 43-70*, Chichester, Wiley-Blackwell.

[FB07]Frowd, C.D., Bruce, V., McIntyre, A. and Hancock, P.J.B. (2007a), “The relative importance of external and internal features of facial composites”, *British Journal of Psychology*, 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. (2007b), “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. (2007c), “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. (2005a), “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. (2005b), "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., Fodarella, 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. (2013a), "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. (2007d), "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. (2012a), "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. (2012b), "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. (2012c), "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. (2012d), "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. (2013b), "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. (unpublished), "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.

Lewis, M.B. (2004), "Face-space-R: towards a unified account of face recognition", *Visual Cognition*, Vol. 11, pp. 29-69.

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

Marsh, J. E., Demaine, J., Bell, R., Skelton, F.C., Frowd, C.D., Röer, J.P. and Buchner, A. (2016), "The impact of irrelevant auditory facial descriptions on memory for target faces: Implications for eyewitness memory", *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., Hancock, P.J.B., Bowie, L., Bruce, V. and Pike, G. (2015), "Are two views better than one? Investigating three-quarter view facial 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.

Skelton, F.C., Frowd, C.D. and Speers, K. (2016), “The benefit of context for facial-composite construction”, *Journal of Forensic Practice*.

Sporer, S.L. and Martschuk, N. (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.

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

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.

Figure 1. Composites constructed in the included studies from (left to right) EvoFIT, Feature and Sketch systems. Composites 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).

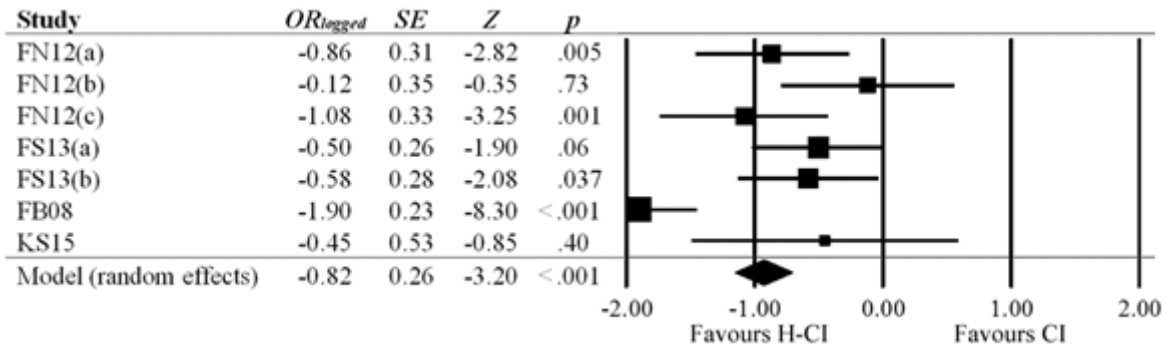

Figure 2. Forest plot of the H-CI versus CI advantage (OR_{logged}) for accurate naming.

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 celebrity 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 well-known famous faces (e.g. David Beckham, Ronan Keating, David Tennant and Prince William).

†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) for the associated Odds Ratio.

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 $R^2 = .17$]. Presented for each predictor is the Beta (B) coefficient (slope of regression line), standard error of *B* (*SE(B)*), Wald (X^2), *DF*, model fit (*p-value*), Odds Ratio effect size (*Exp(|B|)*) and (in square brackets) 95% CI for *Exp(|B|)*. The inequalities under Variable indicate the direction of each difference: *B* values may be positive or negative depending on coding (variables with the lowest numerically coded-category are shown underlined, but see also text). For ease of interpretation, Odds Ratios are shown with values greater than 1.0, rather than allowing them to appear as a Risk Ratio (a value less than 1.0). *N* is the minimum number of comparisons involved in the calculation; for instance, *N* = 7 for Interview as there are 34 conditions for CI and 7 for H-CI.

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*.

Table 4: Accurate and inaccurate naming for Logistic Regression Model B (EvoFIT composites).

Variable	*N*	*B*	*SE(B)*	*X²*	*DF*	*p*	*Exp(|B|)*	
Accurate								
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	
Inaccurate								
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. Accurate Model [$X^2(2) = 104.0$, $p < .001$, Cox and Snell $R^2 = .04$, Nagelkerke $R^2 = .06$]. Inaccurate Model [$X^2(4) = 180.1$, $p < .001$, Cox and Snell $R^2 = .11$, Nagelkerke $R^2 = .15$]. See Table 2, *Note*.

Footnotes

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

[ii] Studies contained a *pre*-screening phase to check that targets were *unfamiliar*: Constructors glanced at a (randomly-selected) target; if the face was reported familiar, another target was presented likewise, and participants encoded the first unfamiliar face. A *post*-screening phase presented target images to 'naming' participants *after* composites had been seen, to check that identities were *familiar*. Also applied was an *a-priori* rule: each participant was required to correctly name most targets (typically $M > 75\%$) for their data to be analysed (if not, another participant was recruited as replacement).

[iii] Models were re-run without backward elimination. While this was not necessary for Model A (accurate), as all variables were reliable, for other models, saturated and final solutions contained the same reliable predictors with virtually identical coefficients.

[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%.