

2 *1. General Method*

3
4 *1.1 Stage 1: Target Encoding (Procedure and Materials, all experiments)*
5

6 Mirroring the forensic situation, participant-witnesses who were unfamiliar with the target-
7 identity pool were recruited to Stage 1 of the experiment. Participants first briefly encoded
8 the face of a single unfamiliar target identity (for 60 seconds in Experiment 1, and for a more
9 ecologically-valid period of 30 seconds in subsequent experiments; Frowd et al., 2015). Faces
10 were viewed under intentional encoding instructions—that is, participants were made aware
11 that they would later construct a composite of the presented face¹. It was important to keep
12 the experimenter, who would later operate the composite system, naïve to the pool of target
13 identities. Firstly, experimenters all reported to be unfamiliar with the relevant target pool
14 from the outset, and secondly, to maintain naivety, the experimenter left the room while the
15 participant either opened and viewed the allocated digital file (Experiment 3) or turned face-
16 up the piece of paper on which the target’s face was printed (all other experiments).

17 To facilitate generalisation of results, different target identity pools were purposely used in
18 each experiment (*see* interim method sections). However, all target photographs were
19 prepared and presented to the same standard across experiments. Specifically, good-quality
20 photographs of each target identity, sourced from the internet, depicted the head and
21 shoulders of the individual, who was adopting a front-facing, neutral pose, with minimal
22 facial hair and no adornments (e.g., no target faces had a nose stud) that might otherwise
23 render the face too distinctive. Per experiment, a copy of these target photographs was
24 prepared in an electronic document for each condition, in colour, at 8 cm width x 10 cm

¹ Eyewitnesses tend to use this type of encoding (Fodarella et al., 2021); indeed, spontaneous sub-vocalisations during encoding (e.g., ‘light eyes, arched eyebrows’) demonstrate an awareness that retrieval of facial detail may be required at a later date.

1 height, one per A4 page. For face-to-face interactions (Experiments 1, 2 and 4), these
2 documents were reproduced using a good quality printer.

3 Identity replacements were made for any participant who reported to be familiar with the first
4 facial identity they were originally asked to encode. This circumstance occurred four times in
5 Experiment 3 and once in Experiment 4, with no replacements made in Experiments 1 and 2.

6 *1.2 Stage 1: Self-administered written interview (Materials and Procedure, Experiments 2 –*
7 *4).*

8 During the target-viewing session, participants assigned to the early recall condition received
9 a sealed envelope from the experimenter (Experiments 2 and 4). They were told to open the
10 envelope 3-4 hours later, and follow the printed instructions therein, which asked them to
11 write down as much as they could remember about the face on the enclosed A4 sheet of paper
12 (i.e., a free-recall attempt). While participants were not subsequently reminded to complete
13 the task, they were requested to return this description to the experimenter when they
14 attended their next experimental session (described below), as a compliance check².

15 Participants were not required to review this description ahead of their next experimental
16 session [comprising the CI, or (the original or modified) H-CI, and composite construction]
17 as research suggests that reviewing a retrieval attempt does not facilitate subsequent recall
18 (e.g., Sauerland et al., 2008; Turtle & Yuille, 1994).

19 The procedure for requesting early recall was adapted to be remote for Experiment 3, due to
20 restrictions imposed by the COVID-19 pandemic. Here, 3-4 hours after encoding, the
21 researcher contacted participants assigned to the early-recall condition by telephone,
22 requesting them to write down a description of the target face once the call had ended. In the

² As the written-recall task was designed to be conducted in the absence of the experimenter, no further compliance checks were carried out for this procedural element of the experiment.

1 following meeting, all participants reported that they had completed the exercise, as
2 requested.

3 1.3 Stage 1: Practitioner-led Cognitive Interview (*Materials and Procedure, all experiments*)

4
5 Participant-witnesses began their final experimental session with completion of a three-stage,
6 face-recall Cognitive Interview (CI), which was conducted online for Experiment 3 (via
7 FaceTime or Skype), and in-person for the other experiments. The experimenter first invited
8 the participant to think back to when the target's face had been seen (i.e., as part of *context*
9 *reinstatement*), and to retrieve a good visual image of the face from memory. Once the
10 participant confirmed that this had been achieved, a free-recall stage followed, during which
11 the participant was invited to verbally recall any and all details they could remember about
12 the face, in their own time and words, without guessing, and without interruption from the
13 experimenter. During participant recall, the experimenter wrote down the provided
14 description on an A4 sheet of paper, with descriptors separated according to the sheet's
15 section headers (i.e., for overall facial characteristics, facial shape, hair, eyebrows, eyes, nose,
16 mouth and ears). A cued-recall stage followed wherein the researcher repeated back,
17 verbatim, details the participant had provided, per section header, and asked the participant
18 whether they could recall anything further about that particular facial region or feature (e.g.,
19 'You mentioned to me that the hair was brown and short. Is there anything else you can
20 remember about this feature?'). This cued-recall stage was omitted for participant-witnesses
21 in Experiment 4, as this mnemonic does not appear to facilitate EvoFIT construction (e.g.,
22 Frowd et al., 2015).

23

24

1

2 *1.4.1 Stage 1: Practitioner-Led Holistic-Cognitive interview (Materials and Procedure,*
3 *Experiments 2 and 4)*

4 In addition to the face-recall CI, participant-witnesses in specific conditions of Experiments 2
5 and 4 then immediately completed holistic recall, as part of a Holistic-Cognitive Interview
6 (H-CI), which they were informed would later help them to construct an identifiable image
7 (e.g., Frowd et al., 2012). Here, these participants were asked to reflect silently on the
8 perceived personality of the face, for which 1-minute was given. Next, they were asked to
9 provide seven ratings, anchored on a three-point scale (*low, medium and high*) to reflect how
10 they perceived the face, as a whole, to convey specific personality characteristics. The
11 characteristics (intelligence, friendliness, kindness, selfishness, arrogance, distinctiveness and
12 aggressiveness) were stated aloud sequentially by the experimenter, with the experimenter
13 recording the rating that the participant gave to each prompt. These ratings were recorded on
14 the same sheet that had been used to collect the participant-witnesses CI description.

15 *1.4.2 Stage 1: Practitioner-Led modified eye-region H-CI (Materials and Procedure,*
16 *Experiment 4, only)*

17 In Experiment 4, a third of participant-witnesses were assigned to receive a revised version of
18 the H-CI. For EvoFIT, Skelton et al. (2020) found enhanced composite effectiveness when
19 participants provided the aforementioned holistic ratings twice: once for the whole-face and
20 then again when focusing on the eye region (the area including the eyes and eyebrows).
21 Potentially harnessing Transfer Appropriate Processing (TAP; Morris et al., 1977)
22 mechanisms, this restricted focus aligns with that instructed during EvoFIT array
23 presentation, where witnesses are encouraged to focus on the likeness of the eye-region when
24 making their face selections (Fodarella et al., 2017). Here then, participant-witnesses used the

1 same three-point scale to rate the extent to which they perceived the eye region to convey the
2 same seven characteristics (as above) of the target's character, with the experimenter again
3 recording these ratings on the aforementioned response sheet.

4 *1.5.1 Stage 1: PRO-fit Construction (Procedure, Experiments 1 and 2)*

5 Immediately following the CI (Experiment 1), or H-CI (Experiment 2), participant-witnesses
6 engaged in experimenter-led PRO-fit construction. The experimenter was extensively trained
7 in construction techniques and naïve to the to-be-constructed target identity. The procedure
8 for face construction using PRO-fit is thoroughly described elsewhere (e.g., *see* Fodarella et
9 al., 2015), and so an outline is provided here.

10 The experimenter first independently entered the descriptors provided by the participant-
11 witness during the CI, as recorded on the description sheet, to locate approximately 20
12 'matching' system-housed photographic exemplars, per facial feature (e.g., for the eyes, nose,
13 mouth, etc.). The experimenter then showed the participant the returned exemplar sub-set,
14 per feature, embedded within the context of a whole-face, and the participant was asked to
15 direct the experimenter toward the single best exemplar, per feature category. With these best
16 feature exemplars in place, the participant was then invited to suggest how the likeness of the
17 face could be improved, with the experimenter using editing tools to re-position, re-size and
18 re-shade facial features, as requested. PRO-fit construction took approximately 1-hour,
19 including debriefing.

20

21 *1.5.2 Stage 1: Sketch Composite Construction (Procedure, Experiment 3)*

22 An established procedure of sketch production (e.g., Fodarella et al., 2015; Frowd et al.,
23 2005) was implemented by an extensively-trained, target-naïve, artist. Due to restrictions

1 imposed by the COVID-19 pandemic, interaction with participant-witnesses was carried out
2 via video link (FaceTime or Skype), a procedure previously found to be effective for
3 construction of forensic sketches (Kuivaniemi-Smith et al., 2014). Directly consulting the
4 participant's face description, obtained during the CI, the artist prepared an initial sketch,
5 wherein facial features were faintly drawn. The artist then followed instructions, given by the
6 participant, to improve image likeness, altering feature size, position and shading. Sketched
7 composites took around two hours to construct, including debriefing.

8

9 *1.5.3 Stage 1: EvoFIT Composite Construction (Procedure, Experiment 4)*

10 An extensively-trained, target-naïve experimenter controlled the software. The EvoFIT
11 construction process is described in detail elsewhere (e.g., Fodarella et al., 2015), and thus a
12 brief protocol is presented here. Participant-witnesses first directed the experimenter to a
13 database that matched the previously-seen target for age and gender. Participants were then
14 presented with four screens of 18 'smooth' (texture-averaged) faces that revealed the internal-
15 features region (i.e., the facial area excluding hair, forehead, ears and neck): they were asked
16 to ignore face width but indicate to the experimenter the best two matching items from each
17 of the first three screens, based on the target-likeness of the eye region. The participant-
18 witness could review their selections, and make any replacements, on a fourth screen. This
19 procedure was repeated over four screens of 'textured' faces (presented with variable facial
20 texture), with participants then presented with a combination of previously-chosen smooth
21 and textured faces from which they directed the experimenter towards the single best match.
22 Participants undertook a second experimenter-led iteration, with previous choices combined,
23 to 'evolve' a face. The participant then directed the experimenter to enhance the likeness, first
24 using holistic tools: scales that changed width, weight, age, and 12 further overall properties

1 of the face. The face was then subject to further enhancement: the experimenter could first
2 adjust greyscale shading of features and then feature shape and position on the face. Hair and
3 other external features were added, and the aforementioned software tools were used again,
4 as required, with the aim of creating the best likeness possible. The procedure took
5 approximately 45 minutes, including debriefing.

6 *2.1 Stage 2: Naming (Materials and Procedure, all experiments)*

7 Mirroring the forensic situation, target-familiar participants were recruited to attempt to name
8 the composites produced during Stage 1, with the following procedure conducted in-person
9 for Experiments 1 and 4, and remotely (via FaceTime or Skype) for Experiments 2 and 3.

10 Participant-namers were tested individually, and the task was self-paced. Each participant
11 was randomly allocated to view the composites constructed in only one of the Stage 1
12 conditions of that experiment, with items presented by the experimenter sequentially, in a
13 different random order for each person. Composites were sized to 8 cm (width) x 10 cm
14 (height) in electronic documents. Each document contained 10 composites (one per target
15 identity), each presented individually per A4 page, in greyscale, which were printed to good
16 quality for face-to-face interactions. Participants were asked to name each composite, saying
17 a name if one came to mind; otherwise, a “don’t know” response was acceptable.

18 Responses to composites were scored either as ‘correct’ or ‘incorrect’, with the latter category
19 comprising both “don’t know” responses and mistaken names (i.e., where the participant-
20 namer had offered a legitimate character or actor name that did not match the constructed
21 identity). Response differentiation allowed an assessment of composite effectiveness: while
22 good quality composites attract a high proportion of correct names, composites that are
23 unnamed or frequently attract mistaken names insufficiently resemble target identities, or
24 better resemble another identity, suggesting lower quality.

1 After viewing all composites constructed in their assigned condition, participant-namers were
2 shown photographs of the corresponding target identities to name, to check for suitable
3 familiarity with the target pool. Target photographs were presented sequentially to the
4 participant by the experimenter, were prepared to the same size and standard as composite
5 images, but were shown in colour. Target photographs were presented in a different random
6 order for each person, by identity, and this order differed to the random order of presentation
7 for composite images.

8 As participants were recruited on the basis of being familiar with the target pool, if they
9 failed to recognise either one or two of the identities, data for these associated composites
10 were discarded; if they failed to recognise more, they were replaced by another participant,
11 which happened rarely across the four experiments. The task took around 15 minutes to
12 complete, including debriefing.

13 *3.1 Stage 3: Composite-to-target likeness ratings (Materials and Procedure, all experiments).*

14 Participant-raters tend to judge visual match more harshly for identities with whom they are
15 familiar than unfamiliar (Frowd, 2021), and so target-unfamiliar participants were recruited to
16 Stage 3. As such, data retention principles contrasted with those implemented in Stage 2: if
17 the participant *did* recognise one or two of the target identities (as assessed via a final
18 photograph naming task, described below), their data for those individual composites were
19 discarded; if they recognised more than two, the participant was replaced, with the latter
20 instance occurring rarely across experiments.

21 Participant-raters were tested individually, either face-to-face (Experiments 1 and 4), or
22 remotely (Experiments 2 and 3, via FaceTime or Skype) and the task was self-paced. A
23 within-subjects design was adopted: For each target identity, participants were concurrently
24 presented with all of Stage 1's corresponding composites (i.e., one facial image resulting

1 from each construction condition) and the corresponding target photograph. Composite array-
2 to-target photograph slides were presented randomised by target identity and participant, with
3 both composites and target photograph images sized to the same dimensions as in Stage 2.
4 Per composite-to-target pairing, participants were asked to assess the likeness between the
5 two images, with absolute judgments given in Experiment 1 (i.e., participants made a
6 composite-to-target rating for the first composite and first target identity, before viewing and
7 rating the second composite according to its likeness again to the first target identity, and so
8 on, until they had provided a likeness rating for all composites constructed to resemble that
9 target identity). Subsequent experiments instead required relative likeness judgments to be
10 made (i.e., participants first passively viewed all composites constructed to resemble a
11 particular identity before they sequentially rated the likeness between each of those
12 composites and the same target identity). The latter task variation was made as it can be
13 difficult to judge variation in likeness without first inspecting the relevant composites; a
14 method of presentation that could otherwise produce a range effect (e.g., Poulton, 1975).

15 Across experiments, the likeness rating scale varied: in Experiment 1, ratings were provided
16 on a 15-point scale anchored from ‘*very dissimilar*’ to ‘*very alike*’, while in subsequent
17 experiments, a truncated scale, with better-defined endpoints, was used (i.e., (1 = *very poor*
18 *likeness* ... 7 = *very good likeness*). This decision arose as Experiment 1’s data revealed
19 unequal distribution of ratings across the scale, with participant’s evidencing reluctance to
20 rate with higher scale points (from 8 – 15). For the ensuing GEE and GLMM analyses, this
21 necessitated scale-recoding; specifically scale points of 8 and above were collapsed to a
22 single category (scale point 8) to produce a more equal frequency distribution across the
23 remaining scale points. We hoped to avoid scale recoding in subsequent experiments, as this
24 action reduces the range and veracity of the data. However, participants in Experiments 2 – 3
25 still infrequently selected the highest scale point (of 7) and so similar, although less extreme,

1 value-collapsing was undertaken (i.e., in Experiment 2, scale ratings from 5 – 7 were
2 collapsed to a value of 5; and in Experiment 3, scale ratings of 6 and 7 were collapsed to a
3 value of 6). Dependent on condition assignment, participants in Experiment 4 demonstrated a
4 reluctance to use lower and higher scale points, respectively, thus for all participant responses
5 values from 1 to 3 were recoded as 3, and values 5 to 7 as 5.

6 To assess for suitable levels of target (un)familiarity, participant-namers then viewed each
7 target photograph, sequentially, in a different random order per participant, and attempted to
8 provide a name for each. This task also took around 15 minutes to complete, including
9 debriefing.

10 4.1 *Power and Inferential Analyses*

11 4.1.1 Naming Analyses

12 ***Approach***

13 Generalized Estimating Equations (GEE) were used to analyse participant naming responses
14 to composites for all experiments presented in this paper (SPSS Version 29 using GENLIN,
15 IBM Corp.). This regression technique uses a binary approach to composite naming
16 responses. Two main analyses were conducted, one for *correct* naming (coded as *1* when the
17 given name was accurate, and *0* otherwise) and the other for *mistaken* naming (coded as *1*
18 when the given name was erroneous, and *0* otherwise), with a consideration of both indices
19 affording a comprehensive assessment of composite quality.

20

21 For all experiments, two GEE analyses were first conducted by the second author and
22 checked by the last. The first analysis was *by-participants*, a conventional analysis to assess
23 the extent to which results generalise to other participants. The second, *by-items*, to confirm
24 that results generalise to other stimuli, thus avoiding suggestion of a stimuli-as-a-fixed-effect

1 fallacy (Clark, 1973). These analyses were modelled by specifying the coding for the within-
2 subjects' variable as *items* (identities or stimuli in the experiment) in the former, and
3 *participant-namers* in the latter. Both analyses produced the same pattern of significant and
4 non-significant differences, except for one additional significant difference for (the less
5 forensically-important) mistaken naming measure in the by-items analysis in Experiment 2,
6 and so, for brevity, by-participant analyses are presented in Results, with further details
7 provided in Appendix A, and by-items analyses in Appendix B.

8
9 The statistical analysis as described can be considered good practice when there is need to
10 analyse participant responses from psychological experiments. In addition to participant-
11 namers and items, the current forensic application involved a third source of variation:
12 *participants-witnesses* (i.e., participants who had constructed the composites). The random
13 effect of participant-witnesses increases model complexity markedly, usually impacting
14 statistical power, and was accounted for in a combined measure across experiments. This
15 additional analysis provides a single estimate of the overall size of the effect for the two
16 predictors of interest, *Early Recall* and *Interview Type*.

17
18 In each analysis, similar to Repeated Measures ANOVA, participant responses were modelled
19 as being equally correlated, achieved by selecting an Exchangeable Working Correlation
20 Matrix. Unlike tests such as ANOVA, regression models are usually subject to an iterative
21 process to select predictors. As such, to lessen the chance of making a Type II error,
22 predictors (IVs) were maintained in the model based on the established criteria for regression
23 analyses of $p < .1$ (e.g., Field, 2018). Both Model-based and Robust covariance estimators
24 were conducted, with smaller standard error (*SE*) values for a predictor's coefficient (*B*)
25 indicating a better overall fit of the data. *SE(B)* values emerged much lower for Model-based

1 (cf. Robust), or varied little, and so, as Model-based is available in more statistical packages,
2 this estimator was selected throughout. Further, for all analyses, coefficients, standard errors
3 and confidence intervals were checked for appropriate values, neither too low nor too high,
4 that might otherwise indicate an issue with model fit.

5

6 In terms of reported statistics, we present the results of the analyses comprehensively, as is
7 best practice (e.g., Bolker et al., 2009). However, one common statistic not reported is the
8 inferential fit for a model's intercept (i.e., to test the null hypothesis that the fixed intercept,
9 B_0 , equals 0). For the research, this inferential statistic is not necessary (but could be derived
10 from the given values) and so, for brevity, only B and $SE(B)$ are reported for the intercept.

11

12 Using the above approach, we also took this opportunity to conduct analyses using GLMM
13 (Appendices C-E). This regression approach involves fixed effects (predictors, or IVs, as
14 modelled by GEE), but also random effects (e.g., the influence of participants and stimuli).
15 As such, it provides a combined by-participants and by-items model that is gaining popularity
16 (Meteyard & Davies, 2020). At the time of writing, GLMM only seems to have been used to
17 formally analyse responses to composites in one prior publication (Erickson et al., 2022), and
18 so we compared the established GEE method with GLMM to provide evidence for or against
19 the applicability of the latter technique.

20

21 ***Statistical Power***

22 A between-subjects design was followed for face construction (Stage 1) and composite
23 naming (Stage 2), with appropriate Generalized Estimating Equations (GEE) analyses
24 planned. To be of practical significance, at least a medium effect size was desired. Previous,
25 similar work (e.g., Erickson et al., 2022; Frowd et al., 2013; Martin et al., 2017; Portch et al.,

1 2017; Skelton et al., 2020) indicated that a minimum of 10 participants per condition was
2 required for face construction and composite naming, respectively, with the appropriateness
3 of these estimates assessed by computer simulation.

4

5 Here, participant-namer responses were simulated for each experiment, and then analysed in
6 the same way, using GEE. The same as in the experiments, GEE used a logistic link function
7 to model the dichotomous nature of the DV, and all predictors were coded as nominal
8 variables. As participants attempted to name multiple composites, responses to these images
9 were modelled as being equally correlated by specifying an Exchangeable Working
10 Correlation Matrix. Each set of simulations was repeated 100 times, by-participants and by-
11 items, with the frequency that results emerged significant (i.e., given $p < .05$) reported as a
12 measure of statistical power.

13

14 In Experiment 1, there was one predictor, *Retention Interval*, with four delay intervals
15 (immediate, 3-4 hours, 2 days and 1 week). This variable was modelled as described in
16 Equation 1:

17

18 Equation 1 - Model for a single Predictor in the Regression Equation for Experiment 1:

19

$$20 Y_{ij} = B_0 + (x_{11} * B_{11}) + (x_{12} * B_{12}) + (x_{13} * B_{13}) + (x_{14} * B_{14}) + e_{ij}$$

21

22 Where $x_{11} - x_{14}$ are levels of the predictor *Retention Interval* with associated Beta values (B_{11}
23 to B_{14}). B_0 is the model's intercept. The term e_{ij} is the residual error. For analysis of nominal
24 responses, the equation was subject to the Sigmoidal function, $Y'_{ij} = \text{Exp} (Y_{ij}) / (1 + \text{Exp} ($
25 $Y_{ij}))$.

26

27 Baseline performance was defined relative to immediate construction for an expected mean
28 correct naming of 30% for a computerised feature system (Frowd et al., 2015). It was realised

1 for the model's Constant (B_0) by random sampling of a Normal distribution based on a value
2 of -0.85, with SD set to 0.1 to give a sensible range ($\pm 2 SD$) from 26 to 34% between
3 participant-namers. Based on expectation, $Exp(B)$ was modelled to *reduce* naming
4 successively by a medium effect across each delay interval (sampling B from a random
5 Normal distribution with mean values of -0.92, -1.83 and -2.75, respectively), again with SD
6 = 0.1, to provide variability in participant-namer responses. Residual errors (e_{ij}) were added
7 to each participant-namer response, again using a random Normal distribution ($M = 0.0$), SD
8 = 0.5, again to provide suitably variable individual responses. Finally, as target identities are
9 sometimes not correctly named (typically 1 in 20), we modelled this situation, since
10 associated composite responses cannot be correct and so are removed prior to analyses—a
11 procedure that increases $SE(B)$ and impacts statistical power. Accordingly, 5% of cases were
12 selected by chance to be an unfamiliar identity and then processed accordingly. Simulation
13 included three random effects: stimulus items (coded 1-10), participant-witnesses (1-40), and
14 participant-namers (1-40).

15

16 *Retention Interval* was significant as a main effect (i.e., with an omnibus value of $p < .05$) for
17 each simulation, by-participants and by-items. Reverse Helmert contrasts emerged significant
18 the vast majority of the time, 91% by-participants and 94% by-items. Power was weakest for
19 the first contrast (i.e., 3-4 hr vs. immediate) and was significant 76% of the time by-
20 participants and 84% by-items; other contrasts were significant over 99%. These simulations
21 indicate that good statistical power has been achieved.

22

23 Experiment 2 involved a factorial design with predictors of *Early Recall* and *Interview Type*
24 (see Equation 2, below). Computer simulation was based on a medium, positive, additive
25 effect for these two predictors (i.e., $Exp(B) = +2.5$) using the proposed design (e.g., 10

1 different stimuli items, and 10 participants / group for both participant-witnesses and
2 participant-namers). Baseline performance for PRO-fit was taken from Experiment 1 at the
3 two-day delay interval, a mean of 9% correct (i.e., $B_0 = -2.31$); other parameters were the
4 same as described for Experiment 1, above (e.g., same settings for SD). Simulation by-
5 participants and by-items revealed that these two predictors were significant between 95 and
6 97% of the time, again indicating good statistical power.

7
8

9 Equation 2 - Model for each Predictor in the Regression Equation for Experiment 2:

10

$$11 \quad Y_{ij} = B_0 + (x_1 * B_1) + (x_2 * B_2) + e_{ij}$$

12

13 Where x_1 is the predictor for *Early Recall* and x_2 for *Interview Type* with associated Beta
14 values (B_1 and B_2). See Equation 1 for definition of other terms. Note that terms for an
15 interaction were not included since effects were predicted to be additive.

16 Experiment 3 involved a single factor, *Early Recall*. Relative to computerised feature
17 systems, composites from Sketch are usually constructed more effectively at a long retention
18 interval (e.g., $M = \sim 15\%$ in Frowd et al., 2015, and $\sim 35 - 45\%$ in Kuivaniemi-Smith, 2023),
19 and so a medial baseline of 30% correct was specified, giving $B_0 = -0.85$. Using other settings
20 from the first simulation and modelling a medium effect for the predictor, *Early Recall*, this
21 fixed effect was significant 83% by-participants and 84% by-items, once again indicating
22 good statistical power.

23 Equation 3 - Model for each Predictor in the Regression Equation for Experiment 3:

24

$$25 \quad Y_{ij} = B_0 + (x_1 * B_1) + e_{ij}$$

26 Where x_1 is the predictor for *Early Recall* with associated Beta value (B_1). (See Equation 1
27 for definition of other terms.)

1 Experiment 4 involved a single factor, *Interview Type*, comprising three levels, Level 1 (CI),
2 Level 2 (H-CI) and Level 3 (Early Recall plus CI) (see Equation 4, below). Baseline naming
3 is usually higher for this type of composite system, and here performance was set to 45%
4 correct based on Frowd et al. (2012), giving $B_0 = -0.20$. We again modelled a medium effect
5 from Level 1 to 2, and then again from Level 2 to 3. Other settings were the same as in the
6 previous simulations. *Interview Type* emerged significant each run, by-participants and by-
7 items. Post hoc tests (comparing Levels 1, 2 and 3) were conducted using Parameter
8 Estimates. Level 2 emerged significantly greater than Level 1 on 85% of occasions by-
9 participants and 88% by-items; Level 3 was greater than Level 1 on every occasion. Re-
10 running the analyses with a different sorting order specified for target and predictors, to
11 obtain parameter estimates for Level 3 versus Level 2, revealed that this third contrast was
12 significant 75% of the time by-participants and 77% by-items. Simulations thus indicated
13 good statistical power.

14
15 Equation 4 - Model for each Predictor in the Regression Equation for Experiment 4:

16
17
$$Y_{ij} = B_0 + (x_{i1} * B_{11}) + (x_{i2} * B_{12}) + e_{ij}$$

18 Where x_{i1} and x_{i2} are levels of the predictor *Interview Type* for H-CI and H-CI plus early
19 recall, with associated Beta values (B_{11} and B_{12}). See Equation 1 for definition of other terms.

20
21 So, overall, while the estimated sample sizes may seem small, they have been successfully
22 used in previous research (e.g., see references above), and are here supported by simulation.
23 Indeed, this sample size was able to reliably detect a medium effect in each of the
24 experiments reported in this paper (see General Discussion and Appendix D). Also, it was
25 sufficient for analysing correct naming responses using a complementary regression
26 technique, Generalized Linear Mixed Models (GLMM; see Appendix C).

1 4.1.2 Likeness Ratings

2 Prior studies using a similar design (within-subjects, identity blocked by target) and GEE for
3 analysis, have recruited between 12 and 30 participant-raters (e.g., Brown et al., 2020;
4 Richardson et al., 2020; Skelton et al., 2020), with a small effect detected ($Exp(B) \geq 1.5$). We
5 followed these extant sample sizes, recruiting between 15 and 18 participant-raters, per
6 experiment.

7

8 GEE (SPSS Version 29 using GENLIN, IBM Corp.) were also used to analyse participant-
9 rater responses for the ordinal-level ratings of composite likeness. We followed the approach
10 outlined for analysing naming responses, above (Section 4.1.1).

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

References

- 1
- 2 Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H.,
3 & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for
4 ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.
- 5 Brown, C., Portch, E., Nelson, L., & Frowd, C. D. (2020). Reevaluating the role of
6 verbalization of faces for composite production: Descriptions of offenders
7 matter! *Journal of Experimental Psychology: Applied*, *26*, 248–265.
- 8 Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in
9 psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–
10 359.
- 11 Erickson, W. B., Brown, C., Portch E., Lampinen, J. M., Marsh, J. E., Fodarella, C., Petkovic,
12 A., Coultas, C., Newby, A., Date, L., Hancock, P. J. B., & Frowd, C. D. (2022). The
13 impact of weapons and unusual objects on the construction of facial composites,
14 *Psychology, Crime & Law*, *30*(3), 207–228.
- 15 Field, A. (2018). *Discovering statistics using SPSS*. 5th Ed. Sage: London.
- 16 Fodarella, C., Frowd, C. D., Warwick, K., Hepton, G., Stone, K., Date, L., & Heard, P.
17 (2017). Adjusting the focus of attention: helping witnesses to evolve a more
18 identifiable composite. *Forensic Research & Criminology International*, *5*(1), 00143
- 19 Fodarella, C., Kuivaniemi-Smith, H. J., Gawrylowicz, J., & Frowd, C. D. (2015). Forensic
20 procedures for facial-composite construction. *Journal of Forensic Practice*, *17*, 259–
21 270.
- 22 Fodarella, C., Marsh, J. E., Chu, S., Athwal-Kooner, P., Jones, H. S., Skelton, F. C., Wood, E.
23 Jackson, E., & Frowd, C. D. (2021). The importance of detailed context reinstatement
24 for the production of identifiable composite faces from memory. *Visual Cognition*,
25 *29*(3), 180–200.
- 26 Frowd, C. D. (2021). Forensic Facial Composites. In Toglia, M., Smith, A., & Lampinen, J.
27 M. (Eds.) *Methods, Measures, and Theories in Forensic Facial-Recognition* (pp. 34–
28 64). Taylor and Francis: UK.

- 1 Frowd, C. D., Carson, D., Ness, H., Richardson, J., Morrison, L., McLanaghan, S., &
2 Hancock, P. J. B. (2005). A forensically valid comparison of facial composite systems.
3 *Psychology, Crime & Law, 11*, 33–52.
- 4 Frowd, C. D., Erickson, W. B., Lampinen, J. L., Skelton, F. C., McIntyre, A. H., & Hancock,
5 P. J. B. (2015). A decade of evolving composite techniques: regression- and meta-
6 analysis. *Journal of Forensic Practice, 17*, 319–334.
- 7 Frowd, C. D., Nelson, L., Skelton F. C., Noyce, R., Atkins, R., Heard, P., Morgan, D., Fields,
8 S., Henry, J., McIntyre, A., & Hancock, P. J. B. (2012a). Interviewing techniques for
9 Darwinian facial composite systems. *Applied Cognitive Psychology, 26*, 576–584.
- 10 Frowd, C. D., Skelton, F. C., Hepton, G., Holden, L., Minahil, S., Pitchford, M., McIntyre,
11 A., Brown, C., & Hancock, P. J. B. (2013). Whole-face procedures for recovering
12 facial images from memory. *Science & Justice, 53*(2), 89–97.
- 13 IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 29.0. Armonk, NY:
14 IBM Corp
- 15 Kuivaniemi-Smith, H. J. (2023). Understanding and improving the effectiveness of sketch
16 facial composites. [PhD Thesis]. University of Lancashire, UK.
- 17 Kuivaniemi-Smith, H. J., Nash, R. A., Brodie, E. R., Mahoney, G., & Rynn, C. (2014).
18 Producing facial composite sketches in remote cognitive interviews: A preliminary
19 investigation. *Psychology, Crime & Law, 20*, 389–406.
- 20 Martin, A. J., Hancock, P. J. B., & Frowd, C. D. (2017). Breath, relax and remember: an
21 investigation into how focused breathing can improve identification of EvoFIT facial
22 composites. In *Proceedings of the 2017 Seventh International Conference on*
23 *Emerging Security Technologies (EST)* (pp. 79–84). Institute of Electrical and
24 Electronics Engineers.
- 25 Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects
26 models in psychological science. *Journal of Memory and Language, 112*, 104092.

- 1 Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer
2 appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–
3 533.
- 4 Portch, E., Logan, K., & Frowd, C. D. (2017). Interviewing and visualisation techniques:
5 attempting to further improve EvoFIT facial composites. In *Proceedings of the 2017*
6 *Seventh International Conference on Emerging Security Technologies (EST)* (pp. 97–
7 102). Institute of Electrical and Electronics Engineers.
- 8 Poulton, E. C. (1975). Range effects in experiments on people. *American Journal of*
9 *Psychology*, *88*, 3–32.
- 10 Richardson, B. H., Brown, C., Heard, P., Pitchford, M., Portch, E., Lander, K., Marsh, J. E.,
11 Bell, R., Fodarella, C., Taylor, S. A., Worthington, M., Ellison, L., Charters, P., Green,
12 D., Minahil, S., & Frowd, C. D. (2020). The advantage of low and medium
13 attractiveness for facial composite production from modern forensic systems. *Journal*
14 *of Applied Research in Memory and Cognition*, *9*(3), 381–395.
- 15 Sauerland, M., Holub, F., & Sporer, S. (2008). Person descriptions and person identifications:
16 Verbal overshadowing or recognition criterion shift? *European Journal of Cognitive*
17 *Psychology*, *20*, 497–528.
- 18 Skelton, F. C., Frowd, C. D., Hancock, P. J. B., Jones, H. S., Jones, B. C., Fodarella, C.,
19 Battersby, K., & Logan, K. (2020). Constructing identifiable composite faces: The
20 importance of cognitive alignment of interview and construction procedure. *Journal*
21 *of Experimental Psychology: Applied*, *26*, 507–521.
- 22 Turtle, J. W., & Yuille, J. C. (1994). Lost but not forgotten details: Repeated eyewitness recall
23 leads to reminiscence but not hypermnesia. *Journal of Applied Psychology*, *79*(2),
24 260–271.

25

26

27