
Article:
Portch, E., Brown, C. orcid.org/0000-0001-9697-4878, Fodarella, C. et al. (20 more authors) (Accepted: 2025) The impact of forensic delay: facilitating facial composite construction using an early-recall retrieval technique. Ergonomics. ISSN 0014-0139 (In Press)

https://doi.org/10.1080/00140139.2025.2519876

[Accepted, Uncorrected proof to be published in *Ergonomics*. The final article will be available, upon publication, via its DOI: https://doi.org/10.1080/00140139.2025.2519876]

The impact of forensic delay: facilitating facial composite construction using an early-recall retrieval technique

Emma Portch (1*; ORCID: 0009-0008-7895-8100), Charity Brown (2; ORCID: 0000-0001-9697-4878), Cristina Fodarella (3; ORCID: 0000-0001-5551-3450), Elizabeth Jackson (3), Peter J. B. Hancock (4; ORCID: 0000-0001-6025-7068), Colin G. Tredoux (5; ORCID: 0000-0002-9653-786X), Michael B. Lewis (6; 0000-0002-5735-5318), Chang Hong Liu (1; ORCID: 0000-0002-2426-4014), John E. Marsh (3, 7; ORCID: 0000-0002-9494-1287), William Blake Erickson (8; ORCID: 0000-0002-2765-3699), Nicholas Philip Mitchell (9), Chiara Fasching (9), Linda Tran (3), Ellena Wood (3), Elaine A. Damin (3), Leonie Robertshaw (10), James Michael Lampinen (11; ORCID: 0000-0002-5854-521X), Louisa Date (3), Spike Joyce (3), Leonie Brooks (3), Ariell Farrow (3), Tom Barnes (3) and Charlie D. Frowd (3; ORCID: 0000-0002-5082-1259)

(1) Department of Psychology, Bournemouth University, Poole, BH12 5BB
(2) School of Psychology, University of Leeds, Leeds, LS2 9JT
(3) School of Psychology, University of Lancashire, Preston, PR1 2HE
(4) Psychology, Faculty of Natural Sciences, University of Stirling, Stirling, FK9 4LA
(5) Department of Psychology, University of Cape Town, Rondebosch, 7701
(6) School of Psychology, Cardiff University, Cardiff, CF10 3AT
(7) Department of Health, Learning and Technology, Luleå University of Technology, Luleå, Sweden, 971 87
(8) Department of Life Sciences, Texas A&M University, San Antonio, TX
(9) School of Engineering and Computing, University of Lancashire, Preston, PR1 2HE
(10) School of Life Sciences, University of Dundee, Dundee, DD1 5EH
(11) Psychological Science University of Arkansas, Arkansas, Fayetteville, AR 72701

*Corresponding author: Emma Portch, Department of Psychology, Bournemouth University, Poole BH12 5BB, eportch@bournemouth.ac.uk

Word Count (exc. Abstract, table content / titles, figure captions, references, and appendices): 7,929

Memory for facial features deteriorates over time, diminishing one's ability to construct an accurate visual likeness of a face (i.e., a facial composite). In Experiment 1, we investigated how delay affects composite construction. Participants recalled an unfamiliar face during a Cognitive Interview (CI) and constructed a feature composite across four post-encoding retention intervals. Correct composite naming declined sharply after a 3–4 hour retention interval, remained stable at two days, and dropped to floor-level after one week. Experiments 2–4 examined how composite effectiveness was influenced by the incorporation of two factors: (a) a novel, self-administered written face-recall attempt, conducted 3-4 hours after encoding, and (b) a standard or modified holistic recall elicited immediately before construction. Participant-witnesses created more identifiable likenesses when early recall was invited, suggesting that this intervention consolidated and enhanced access to facial-feature information. The addition of a character-based interview further improved both feature and holistic composites.

(150 words)

Keywords: facial composite; face memory; retention interval; self-administered interview; testing effect

Practitioner Summary: We identify two simple, practical techniques to improve the effectiveness of facial composites across different systems. Firstly, eliciting written descriptions of the face from witnesses, shortly after encoding. Secondly, asking witnesses to rate how they perceive aspects of the target's personality from their face (holistic recall) immediately before construction.

(49 words)

Introduction

Facial composites are visual likenesses, typically created during forensic investigations by witnesses and / or victims of crime, to resemble offenders with whom they were previously unfamiliar. The resulting image is usually circulated within a police force, or more widely, to prompt identification by someone who is familiar with the face. The police may also compare composites to mugshot images of potential suspects to assess possible identity matches.

Composites were originally hand-sketched by forensic artists, but two types of computerised system were later developed to allow face construction by interviewers without artistic training. Firstly, feature systems (e.g., E-FIT and PRO-fit in Europe and FACES and Identikit 2000 in the US) require a witness to select individual facial features (i.e., eyes, noses and mouths) from large photographic databases, which are then edited to enhance their resemblance to the offender's. Secondly, better-performing modern holistic interfaces (e.g., EvoFIT; ID; EFIT-V / 6; Frowd, 2021; Solomon & Gibson, 2013; Tredoux et al., 1999) prompt witnesses to select whole-face images from multi-face arrays that best resemble the offender. These selections are then combined to evolve a likeness, whose similarity can be enhanced via further editing.

Accurate construction relies on witnesses' ability to access their memory trace for the face. Prior to construction, witnesses typically relay and consolidate facial detail via completion of a specifically-modified Cognitive Interview (CI; Fisher et al., 1987), which continues to encourage both quality and quantity of report, while omitting some mnemonics present in

event-recall interviews (Ashkenazi & Fisher, 2022). Following rapport building, witnesses are instructed to freely-recall a description of the face, with further information sometimes probed using repeated free-recall attempts, or via interviewer-led cued-recall prompts (Frowd, 2023). For feature and hand-sketched composites, interview-obtained facial descriptors are used directly by practitioners to inform initial feature selection, or to roughly draw feature-shapes (e.g., Frowd et al., 2005c, 2012a). Across all systems (including holistic interfaces), witnesses must later consult their own, retained memory for these details to improve likeness by directing edits to feature shape, size and contrast (Frowd et al., 2012a).

Since face memory influences composite effectiveness, it is predicted that variables which negatively influence one may similarly impact the other. Research demonstrates that interview-elicited face recall rapidly declines as the post-encoding retention interval increases (e.g., Ebbesen & Rienick, 1998; Ellis et al., 1980). Consistently, composite effectiveness also reduces markedly with increasing retention interval. Feature composites are poorly recognised after a forensically-typical retention interval of 1-2 days (~ 10-20% lower than those constructed immediately; e.g., Frowd et al., 2005b, 2005c, 2007).
Decrements by retention interval are also observed for sketch and holistic systems (Frowd et al., 2015).

Developing a new approach for forensic practitioners

When construction occurs after a forensically-typical retention interval, preserving access to face memory through recall-consolidation techniques may be expected to improve composite effectiveness. Supporting this, Brown et al. (2017) found that PRO-fit composites were better recognised when participant-witnesses verbally described the face to a practitioner during

two Cognitive Interviews (CIs): one conducted 3-4 hours after face encoding, and another one day later, immediately before construction. This was superior to using a single CI conducted at either the early or late interval. These findings may reflect a testing effect (Roediger & Karpicke, 2006), a concept inherent within the Transfer Appropriate Processing (TAP) framework (Morris et al., 1977). Here final performance on a cognitive task (composite construction) can be improved when its completion is preceded by congruent activities that prime and additively strengthen the same, required cognitive set (Adesope et al., 2017; Yang et al., 2021). Applying this theory to Brown et al.'s (2017) findings, elicitation of early recall first likely prevents decay of face memory (Ellis et al., 1980). Repeated recall then consolidates the memory trace (*see* Odinot et al., 2013 for similar event-recall findings), enhancing the success of each subsequent recall attempt (Roediger & Karpicke, 2006), despite their progressively increased temporal displacement from face encoding (Whitten & Bjork, 1977). This additively-strengthened memory trace may aid the witness to: choose between exemplars within a feature system, direct an artist to sketch feature shapes, and make later feature-based enhancements across all interfaces (Frowd et al., 2005c, 2012a).

Both for practical and theoretical reasons, the present work explores whether similar benefits arise when early recall is instead self-administered and written (cf. verbal and practitioner-led). Firstly, self-administered interviews require no oversight from a practitioner; by placing lesser demands on police resources, a witness may conduct the procedure as soon as possible after the crime (Gabbert et al., 2009), before marked face-memory decay (Ellis et al., 1980). Secondly, while some findings suggest that elicitation of feature-based verbal descriptions can instate a processing style that hinders later attempts to recognise a face holistically – a verbal overshadowing effect (e.g., Schooler & Engstler-Schooler, 1990) – description

accuracy may be an important mediator of such effects (Meissner et al., 2001). By this account, feature-based recall attempts are thought to overwrite the originally-encoded visual face memory trace, thus highly-accurate verbal templates may aid composite construction (e.g., Meissner et al., 2001). Indeed, Brown et al. (2020) found a positive relationship between verbal description accuracy and subsequent PRO-fit composite effectiveness, an effect that decreased with insertion of a post-description delay, which presumably weakened access to a useful verbal template. Written recall is negotiated more slowly and effortfully than verbal recall (e.g., Kellogg, 2007), and thus it may afford further opportunity for witnesses to carefully monitor the accuracy of their report, omitting or editing information about which they are less sure (e.g., Sauerland et al., 2014; Sauerland & Sporer, 2011). In support, Miura and Matsuo (2021) found that comparatively more accurate event- and person-related details were given in an interim interview, conducted between encoding and a later interview, when recall was written versus verbal (though *see* Sauerland & Sporer, 2011). Preliminary work also supports the utility of written, early recall for EvoFIT accuracy (Damin, 2018): Relative to when EvoFITs were produced after a verbal, practitioner-led CI immediately before construction, composite quality was improved when a self-administered, early written-recall was added to this protocol, with these same benefits unreplicated when early recall was negotiated verbally. While Damin (2018) did not directly compare differences in description accuracy according to early-recall modality, the above mechanisms may be implicated.

Both effective recall *and* recognition support face construction. According to a TAP framework (Morris et al., 1977), composite effectiveness may be further enhanced if construction is preceded by multiple tasks that respectively prime separable recall (early and repeated interviewing) and recognition processes (e.g., Shapiro & Penrod, 1986). A candidate

for priming recognition mechanisms is the Holistic-Cognitive Interview (H-CI; Frowd et al., 2008), wherein witnesses are asked to consult the offender's whole-face image, held in memory, and rate how it conveys specific aspects of that individual's personality (e.g., extroversion). When used independently of early and repeated feature-based recall, and positioned between a standard CI and construction, the H-CI improves the effectiveness of PRO-fit and EvoFIT, but not sketch, composites (Frowd et al., 2008, 2015; Skelton et al., 2020). As the H-CI consistently appears in practitioner protocols (e.g., Frowd et al., 2019; Solomon & Gibson, 2013), it is of practical importance to assess whether it retains its utility when combined with novel interventions (early and repeated recall). Recall and recognition mechanisms are differentially important for supporting procedural stages under different construction systems, and thus we make system-specific predictions for the likely independent benefits afforded by each technique when they are employed together (*see* interim introductions for Experiments 2 – 4).

Across our experiments, we first confirm that composite effectiveness diminishes as face memory declines over increasing retention intervals. We then examine whether early written and repeated feature-based recall attempts can counteract memory loss and improve the effectiveness of PRO-fit, sketch, and EvoFIT composites when created after a typical forensic delay. For PRO-fit and EvoFIT, we also investigate whether these novel interventions work synergistically with an established interviewing protocol (the H-CI), which (by virtue of character attribution) is hypothesized to enhance recognition, rather than recall, mechanisms.

EXPERIMENTS 1–4: GENERAL APPROACH

METHOD

A three-stage procedure was employed for all experiments. Unique, hypothesis-naïve participants were opportunity-sampled, per experiment, and were separately recruited to each stage: participant-witnesses recalled a single target face and constructed a composite of it, participant-namers attempted to identify the constructed identities by name, and participant-raters assessed the likeness between each composite and its corresponding target face. University-based ethical approval was granted for all experiments. Although individual differences – in age range, gender, locality and sample composition e.g., university staff-to-student ratio – have minimal impact on construction and naming outcomes (e.g., Frowd et al., 2015), appropriate randomisation techniques were applied to mitigate possible impacts: for assignment of participant-witnesses and participant-namers to condition, and for participant-witnesses to target identity. As the materials and procedure largely replicate those reported in our previous work, we provide only a procedural flow-chart here (see Table 1), with further detail available in interim method sections and online Supplementary Materials**.**

[Table 1]

EXPERIMENT 1: PRO-FIT FACE CONSTRUCTION BY RETENTION INTERVAL

Introduction

Experiment 1 compared composite effectiveness when face construction occurred immediately, and at three forensically-relevant retention intervals: 3-4 hours was chosen as the shortest likely interval (Frowd et al., 2005b); two days was chosen as more typical (Frowd et al., 2005a); and one week was also used, as this interval often occurs in serious criminal cases (Frowd et al., 2012b). While PRO-fit was used here for construction, the predicted detrimental impacts of increased retention interval should be common to sketch and EvoFIT (Frowd et al., 2015).

METHOD

Participants

Sample sizes were determined through known practice and verified through computer simulation (see Supplementary Materials Section 4.1.1).

Construction: Participant-witnesses were 40 students at the University of Lancashire, UK (26 female, 14 male; M_{age} = 20.5, *SD* = 1.4, *Range* = 18 – 26 years), compensated with course credit. To reflect forensic practice, participant-witnesses were recruited to be *unfamiliar* with International-level UK footballers (i.e., the pool from which target identities were drawn for this experiment). We hereafter use the term 'target-unfamiliar' to refer to these circumstances.

Naming: Participant-namers were 40 staff and students at the University of Lancashire, UK (33 male, 7 female; M_{age} = 21.6, *SD* = 6.5, *Range*: 17 – 59 years). Aligned again with forensic practice, participant-namers were recruited to be *familiar* with International-level UK footballers (i.e., they were 'target-familiar').

Likeness: Participant-raters were 15 staff and students at the University of Lancashire, UK (11 female, 4 male; M_{age} = 27.3, *SD* = 12.5, *Range* = 18 – 56 years), and all reported being *target-unfamiliar* (*see* Supplementary Materials Section 3.1 for justification).

Apparatus and Stimuli

The targets were International-level UK footballers (see target sizing and presentation information in Supplementary Materials section 1.1). PRO-fit Version 3.5 was used for construction.

Design

The independent variable was *Retention Interval*—the time between viewing the target photograph and recalling and constructing the face. This variable had four levels: immediate, 3-4 hours, 2 days, and 1 week. Ten participant-witnesses and participant-namers were each assigned to one of the four levels of the *Retention Interval* variable, respectively (i.e., a between-subjects design). For composite likeness ratings, the design was within-subjects for *Retention Interval*.

Procedure

In a first session, each participant-witness was shown one target face (see Supplementary Materials Section 1.1). After their assigned retention interval, participants returned individually for a second session where they first completed a three-stage face-recall Cognitive Interview. The interview procedure began with a prompt for the participant to think back to when the target's face had been seen (i.e., as part of *context reinstatement*), and to retrieve a good visual image of that face from memory. Once the participant confirmed that this had been achieved, a *free-recall* stage followed, during which the participant was invited to verbally recall any and all details they could remember about the face, in their own time and words, without guessing, and without interruption from the experimenter. A *cued-recall* stage followed wherein the researcher repeated back, verbatim, details that the participant had freely-recalled for each facial region or feature, and asked the participant whether they could recall anything further (e.g., 'You recalled that the hair was brown and short. Is there anything else you can remember about this feature?'; see Section 1.3 for experiment-specific variations to the Cognitive Interview procedure). Following face-recall, participants engaged in PRO-fit composite constriction (Section 1.5.1). Figure 1 shows examples of composites constructed of the same target identity at each retention interval.

[Figure 1]

RESULTS

Composite naming

Each participant viewed their assigned set of composites, followed by the 10 corresponding target photographs. Target recognition was appropriately high (M = 97.5%). Since composite-namers who failed to identify a target photograph were unlikely to correctly name the corresponding composite, these instances were excluded from analysis (10/400 attempts). Table 2 shows the resulting average correct and mistaken naming rates for composites by *Retention Interval*. As retention interval increased, correct naming and likeness ratings substantially decreased, while mistaken naming showed a tendency to increase.

[Table 2]

Correct Naming: Generalized Estimating Equations (GEE) were used to analyse responses from participant-namers. This technique modelled correct naming scores (*1* = correct, *0* = otherwise) using a *logistic* link function, a *binomial* distribution and an *exchangeable* working correlation matrix to account for non-independence of the 10 responses provided by each participant. In this first experiment, the sole independent variable was *Retention*

Interval, coded as follows: 1 = immediate; 2 = 3-4 hours; 3 = 2 days; 4 = 1 week[1]. The analysis by-participants revealed that the odds of producing a correct name differed across the four levels of *Retention Interval* [$\chi_1^2(3) = 39.13$, $p < .001$]. Post-hoc tests are hindered both when levels of correct naming are low, as observed here at longer retention intervals, and when their proportions are unevenly distributed across conditions. To maintain statistical power, Reverse Helmert contrasts were conducted to provide a trend analysis. These contrasts compare correct naming at each level of a variable to the collated mean across previous level(s). The results illustrate a decline in composite effectiveness: correct naming was worse after (i) 3-4 hours than immediate [*MD* (Mean Difference) = 14.7%, *SE(MD)* = 0.04, $p < .001$], (ii) 2 days than the shorter (immediate and 3-4 hour) intervals [*MD* = 10.7%, *SE(MD)* = 0.03, $p < .001$] and (iii) 1 week than all other (immediate, 3-4 hour and 2 day) intervals [*MD* = 12.9%, *SE(MD)* = 0.02, $p < .001$].

Mistaken Naming: We compared instances where participant-namers provided an incorrect name for the composite (i.e., they had mistaken the composite for a different identity) across the four retention intervals. As before, responses to composites were removed where the corresponding target photograph had not been correctly named. Overall, mistaken names occurred fairly frequently (*N* = 140/390), consistent with previous findings for feature composites (e.g., Frowd et al., 2012c), and peaked at the longest retention interval (Table 2).

GEE analysis revealed different odds of producing a mistaken response by *Retention Interval* [$\chi_1^2(3) = 10.80$, $p = .013$]. Reverse Helmert contrasts showed that composites were

[1] For all experiments, refer to online Supplementary Materials Section 4.1.1 for details of how the GEE models were constructed. Note that, for all analyses, predictors (IVs and their interaction terms) were retained in the model at $\alpha \leq .1$, while $\alpha \leq .05$ was used for subsequent post-hoc and simple-main effects tests (e.g., Field, 2018). Also, see Appendix A for more information regarding the by-participants analyses and Appendix B for complementary analyses by-items; Appendix C presents analyses instead using Generalized Linear Mixed Effects Models (GLMM).

mistakenly named at a higher rate following 1 week compared to (combined) shorter intervals [MD = 23.6%, $SE(MD)$ = 0.06, $p < .001$; Table 2]; other contrasts were non-significant (ps ≥ .43, MD = -0.03 – 0.06).

Composite Likeness Ratings

Mean correct naming of the target photographs was low (7.3%), confirming participant-rater unfamiliarity with the identities (see Supplementary Materials Section 3.1). These 44 cases of correct naming were removed from the analysis, along with 17 cases of erroneous data entry. For the remaining likeness ratings (1 = very dissimilar, 15 = very alike), responses above the scale midpoint of 8 were sparse, and were thus collapsed (recoded as a value of 8) to create eight ordinally spaced categories (see further justification for, and detail of, data recoding procedures in Supplementary Materials Section 3.1). Overall, rated likeness tended to decline as retention interval increased (Table 2).

GEE was used to fit an ordinal logistic regression to participants' rated likeness with a single predictor: *Retention Interval*. The analysis proceeded as before, except for use of (a) an *ordinal* logistic response function, and (b) an ascending order to sort the dependent variable (Rating). The analysis found different odds of rated likeness across *Retention Interval* [$\chi_1^2(3) = 12.15$, $p = .007$] ($\alpha = .1$).

There was a general decline in composite likeness ratings across retention intervals, although effects were weaker than for correct naming: Reverse Helmert Contrasts[2] revealed that likeness ratings of composites constructed after (i) 3-4 hours were marginally lower than

[2] Unlike analyses for naming responses, contrasts were not available in SPSS version 29 when analysing ordinal (rating) data using GEE, and so we conducted three separate models to compute Reverse Helmert contrasts.

immediate [B = -0.32, $SE(B)$ = 0.18, p = .085], (ii) 2 days were equivalent to the shorter (immediate and 3-4 hour) intervals [B = -0.17, $SE(B)$ = 0.16, p = .28], and (iii) 1 week were lower than all other intervals combined [B = -0.41, $SE(B)$ = 0.15, p = .005].

DISCUSSION

Experiment 1 assessed the impact of increasing retention interval on PRO-fit effectiveness when composites were constructed following a single, practitioner-led CI. As hypothesised, memory for the encoded face deteriorated with increasing retention interval (e.g., Ellis et al., 1980), reducing composite effectiveness (Frowd et al., 2015).

As expected, immediately-constructed composites were correctly named most often, with correct naming rates successively decreasing at each longer retention interval. Rated likeness also evidenced a decrease, particularly when comparing the shortest and longest retention intervals. This suggests that composites progressively contained less of the information required for accurate identification. After a period of 1-week, participant-witness memory was so poor that these composites attracted significantly more mistaken than correct names (relative to mistaken naming at all previous, combined delays), indicating that likenesses tended to be too generic and more often resembled other identities.

In real-world settings, composites are typically constructed 1 – 2 days after a crime (Frowd, 2021), and thus we adopt this retention interval in all subsequent experiments. As decline in correct naming of composites from two days relative to immediate construction [$Exp(B)$ = 3.8] represents a medium-to-large effect size (see Table 4, *Note*), techniques that mitigate the documented sharp decline in memory that occurs between these two timepoints (e.g., Ellis et al., 1980) should be particularly valuable. To achieve this aim, we introduce a novel

technique during this retention interval: participant-witnesses were asked to write a detailed face description 3-4 hours after encoding—the shortest timeframe that such an exercise is likely to be feasible in a criminal investigation. This technique should not only protect against loss of face detail from memory (Ellis et al., 1980) but instate a processing style that facilitates witnesses' tasks on the following day (i.e., completion of a second, practitioner-led CI, and selection and editing of facial features during construction).

EXPERIMENT 2: EARLY RECALL FOR PRO-FIT CONSTRUCTION

Introduction

In Experiment 2, we examine whether early written recall can facilitate PRO-fit construction. Echoing our previous arguments, early and repeated recall interventions are expected to most greatly benefit feature composites (cf. other face-production methods), as they bolster both initial feature selection and later feature editing (e.g., Frowd et al., 2005c, 2012a). This novel intervention is implemented alongside the H-CI, an interview-based technique commonly used by practitioners (e.g., Frowd et al., 2019). Theoretically, pre-construction holistic recall should bolster later recognition that a composite has reached a good level of visual likeness. Thus, both early and holistic recall techniques should enhance composite naming through separate, non-interactive mechanisms.

METHOD

Participants

Construction: Participant-witnesses were target-unfamiliar staff and students at the University of Lancashire, UK, and residents of Whitchurch, Shropshire, UK (23 female, 17 male; M_{age} = 27.2, *SD* = 7.6, *Range* = 18 – 49 years). University students received course credit; otherwise, participation was voluntary.

Naming: Target-familiar participant-namers (24 female and 16 male; M_{age} = 27.1, *SD* = 7.2, *Range* = 18 – 49 years) were students at the University of Lancashire, UK. They received course credit for participation.

Likeness Rating: Target-unfamiliar participant-raters were volunteers from Whitchurch, Shropshire, UK (10 male, 8 female; M_{age} = 30.4, *SD* = 8.3, *Range* = 19 – 47 years).

Apparatus and Stimuli

Ten target photographs of current characters from the ITV soap, "Coronation Street" (5 male, 5 female) were used (*see* Supplementary Materials Section 1.1 for further information). The recording sheet for self-administered, written face-recall is described in Section 1.2. PRO-fit Version 3.5 was used for construction.

Design

Construction and Naming: Ten participant-witnesses and participant-namers were each randomly assigned either to construct a single composite, or view the set of composites created, in one of the four conditions determined by the two between-subjects variables: *Early Recall* (early recall or not) and *Interview Type* (CI or H-CI).

Likeness Rating: Eighteen participant-raters assessed likeness for all composites constructed (i.e., *Early Recall* and *Interview Type* were within-subjects).

Procedure

Construction: This part of the procedure was conducted across two sessions. In the first session, all 40 participant-witnesses undertook target encoding (see Supplementary Materials

Section 1.1). At 3-4 hours after target encoding, half of the participants completed a self-administered, written face-recall attempt (Section 1.2), while the other half did not. The second session was scheduled 20-28 hours following target encoding. Here, participant-witnesses either took part in a practitioner-led Cognitive Interview (CI; Section 1.3) or a whole-face Holistic-Cognitive Interview (H-CI; Section 1.4.1). Immediately following the interview, participants completed PRO-fit construction (Section 1.5.1). A total of 40 composites were constructed, 10 per between-subjects condition.

Composite Naming: The procedure was as previously described (see Supplementary Materials Section 2.1).

Likeness Rating: The procedure was the same as that used in Experiment 1, except for (a) use of a condensed rating scale, and (b) a within-item format, where composites and target photographs were presented together (See Supplementary Materials Section 3.1).

RESULTS

Correct Naming

Participant-namers correctly named all target photographs. Mean correct naming of composites (Table 3) was 36.0% (*SD* = 15.2), increasing markedly both for early recall (cf. no early recall) and for H-CI (cf. CI).

[Table 3]

A full-factorial model was used, with *Early Recall* (coded as 0 = no early recall; 1 = early recall) and *Interview Type* (1 = CI; 2 = H-CI) as predictors. GEE found no interaction [p =

.916, 1/*Exp(B)* = 1.03, α = .1][3], and so this term was removed[4]. In the resulting model[5], both individual predictors returned *p*-values that were less than alpha and so were retained (Table 4). In this final model, the odds of a correct response were higher for *Early Recall* [χ_1^2 (1) = 61.51, $p < .001$], following early recall compared to no early recall; and for *Interview Type* [χ_1^2 (1) = 19.36, $p < .001$], for composites constructed following a H-CI rather than a standard face-recall CI.

[Table 4]

Mistaken Naming

Mistaken names were scored as before, occurring in 80/400 responses (*M* = 20.0%). These responses were somewhat lower for composites created following early (cf. no early) recall (*MD* = 7.0%) but somewhat higher after an H-CI (cf. CI) (*MD* = 10.0%). Following the procedure for correct naming, GEE led to removal of the interaction [*p* = .339, *1/Exp(B)* = 1.82] and then *Early Recall* [*p* = .139, *1/Exp(B)* = 1.56], resulting in a model comprising only *Interview Type* [$\chi_1^2(1) = 4.05$, $p = .044$] (Table 5): the odds of eliciting a mistaken response were higher following an H-CI (compared to a CI).

[Table 5]

[3]When less than one, odds ratios [*Exp(B)*] can be difficult to interpret, and so it is advisable to standardise reporting, such as to present the multiplicative inverse, which we have done here, or to reverse the order of categories (Osborne, 2016). Note that the odds ratio can also be expressed by taking the exponential of the absolute value of *B*, *Exp(|B|)*, a format that is convenient for tables (see Appendix E).

[4] When an interaction or an IV is removed from a model, this indicates that the variable does not affect the DV.

[5] Note that, to reduce the chance of making a Type II error, this approach, involving a model containing both predictors, is preferred over the alternative (where a combined model is considered if each predictor is significant in a separate model). For discussion on this issue, see Field (2018), and Reed and Wu (2013).

Composite Likeness Ratings

Participant-raters rarely gave correct names for target photographs ($N = 12$). The analysis followed the same procedure as for naming data, except that an ordinal logistic response function was used and the target was sorted in an ascending order. Despite condensing the rating scale, responses remained sparse at the two highest scale points, necessitating further scale recoding (see Supplementary Materials Section 3.1). Elevated mean likeness ratings (Table 6) highlighted a consistent benefit for early recall (cf. no early recall), while the H-CI only appears to be beneficial (cf. CI) when combined with early recall.

[Table 6]

In a full-factorial model, GEE analysis retained the interaction between *Interview Type* and *Early Recall* [$\chi_1^2 = 17.90$, $p < .001$]. Parameter estimates for this full-factorial model indicated that the odds of rated likeness were higher for composites following: (i) early recall compared to no early recall at each level of *Interview Type* (ps < .006), and (ii) the Holistic-Cognitive Interview (H-CI) compared to Cognitive Interview (CI) with early recall ($p <$ .001), but not without early recall ($p = 1.00$). For brevity, regression coefficients are presented in Table 6, Note.

DISCUSSION

Experiment 2 manipulated *Early Recall* (present or absent) and *Interview Type* (H-CI or standard face-recall CI). Correct naming was significantly higher for composites created after early recall (cf. no early recall) and a H-CI (cf. CI). We therefore replicate the findings of Brown et al., (2017) when using a written, rather than practitioner-led, early recall attempt

(see also Damin, 2018). As anticipated, correct naming rates indicated that *Early Recall* did not interact with *Interview Type*, suggesting that facilitation occurs via the pre-construction priming of separate underlying (recall and recognition) mechanisms. It was also observed that mistaken names significantly increased with H-CI (cf. CI) but reduced for early recall (which was marginally significant by-items). For likeness ratings, an advantage of H-CI (cf. CI) was not observed without early recall. The latter results indicate an unexpected possible interactive, rather than additive benefit, of our manipulations.

EXPERIMENT 3: EARLY RECALL WITH SKETCH FACE CONSTRUCTION

Introduction

Some composites are manually sketched, with a forensic artist creating the composite based on the witness's face description. As with feature construction, we might expect that feature-memory consolidation through early and repeated recall would facilitate both initial guidance of the artist's feature drawings (e.g., Kuivaniemi-Smith, 2023) and subsequent fine-grained refinements of these details. However, the early stages of sketch construction appear to involve more global facial processing than feature construction, as witnesses tend to focus on groups of features rather than individual ones (e.g., Davies & Little, 1990; Laughery et al., 1986). Therefore, the sketch process may benefit less from attempts to enhance recall.

Furthermore, unlike feature construction, sketch-practitioner protocols typically do not include holistic recall, as it has not consistently improved sketch effectiveness (e.g., Frowd et al., 2015). For practical reasons, this experiment therefore focused solely on the potential benefit of early recall.

METHOD

Participants

Construction: Target-unfamiliar participant-witnesses were staff and students at the University of Dundee, UK (17 female, 3 male; M_{age} = 25.2, SD = 9.0, $Range$ = 20 – 62 years).

Naming: Target-familiar participant-namers were 25 staff and students from the University of Dundee (gender and age undisclosed).

Likeness Rating: Target-unfamiliar participant-raters were 18 student volunteers (15 female, 3 male) at the University of Leeds, UK (age information undisclosed).

Apparatus and Stimuli

Photographs of 10 characters (5 female, 5 male) from the UK TV soap "EastEnders", were used. Stimuli were prepared to the previously-described standard (see Supplementary Materials Section 1.1).

Design

Construction and Naming: Ten participant-witnesses were each randomly assigned to produce a single composite with a sketch artist, with or without early recall. A between-subjects design was also implemented at naming: participants-namers either attempted to name composites produced following early recall (n = 13), or without early recall (n = 12).

Likeness Rating: As before, a within-subjects design was employed for this task.

Procedure

Construction: 3-4 hours after target encoding (detailed in Supplementary Materials Section 1.1), half of the participant-witnesses received telephone instructions to complete a self-administered, written face-recall attempt (Section 1.2). Following a 20–28-hour post-

encoding delay, participants engaged remotely in a practitioner-led Cognitive Interview (CI; Section 1.3), immediately followed by sketch construction (Section 1.5.2).

Naming and Likeness: All interactions with participant-namers and participant-raters were conducted via video link, adhering to the procedures previously described in Supplementary Materials Sections 2.1 and 3.1, respectively.

RESULTS

Correct Naming

There were few cases (N = 9/250, M = 3.6%) where a target photograph was not correctly named. Table 7 shows that sketch composites constructed following early recall attracted substantially more correct names compared to those without early recall, with GEE confirming these trends [$\chi_1^2(1) = 4.08$, $p = .043$, Table 8].

[Table 7]

[Table 8]

Mistaken Naming

Mistaken naming occurred much more frequently than for face construction using PRO-fit, at 45.6% overall. However, this rate differed only slightly between early recall (M = 43.8%) and no early recall (M = 47.8%) conditions. Accordingly, GEE indicated that *Early Recall* had no effect on mistaken naming [χ_1^2 (1) = 0.41, $p = .521$, *Exp(B)* = 1.92].

Composite Likeness Ratings

Target photographs were infrequently identified ($N = 2$, $M = 0.6\%$). As before, responses across the highest scale points were sparse, and so data recoding was performed (see Supplementary Materials Section 3.1). GEE indicated an increase in odds of rated likeness following early recall [χ_1^2 (1) = 15.93, $p < .001$] (see Table 9, *Note*).

[Table 9]

DISCUSSION

The experiment assessed the impact of early recall on the effectiveness of sketch composites. In this iteration, early recall was initiated by a phone call (cf. written instructions). Early recall (cf. no early recall) was again beneficial, this time enabling participant-witnesses to construct a sketch composite that was correctly named significantly more often. We assess the comparative strength of recall-based facilitation for feature and sketch systems in the General Discussion.

EXPERIMENT 4: EARLY RECALL FOR EVOFIT CONSTRUCTION

Introduction

This final experiment assesses whether early and holistic recall will improve EvoFIT effectiveness. Here we anticipated an interaction between our two manipulations: the benefit of early written recall might only be realised when the second practitioner-led CI is followed by holistic recall; a prediction that led us to omit an ‘early recall only’ condition from this experiment.

To explain, while the pre-construction priming of recognition mechanisms via holistic recall might enhance witnesses' ability to assess composite resemblance, irrespective of construction system, recognition processes are also crucial for the initial stage of EvoFIT construction (e.g., Frowd et al., 2012a). Here, witnesses must select whole-face images (or facial regions) that best resemble the offender from multi-face arrays. This activity might be hindered by early and repeated feature-based recall. Indeed, while the latter techniques may consolidate a memory trace for later refinement of feature-based details (Brown et al., 2017), they may also lead witnesses to enter the construction phase with a temporary feature-based processing style, suboptimal for whole-face judgments (i.e., a verbal overshadowing effect; Brown & Lloyd-Jones, 2002; Frowd & Fields, 2011; MacLin, 2002; Schooler & Engstler-Schooler, 1990). Positioning holistic recall between face recall and construction should temporarily release witnesses from this processing style, enabling them to utilise recognition mechanisms more effectively in the initial stages of EvoFIT construction, while leaving a recall-consolidated feature memory trace spared for later consultation.

To further prime recognition mechanisms and align witness processing across construction stages, an additional TAP-informed manipulation was included in this experiment (i.e., Skelton et al., 2020). After participant-witnesses gave holistic ratings to the entire target face, as they had done in Experiment 2, they were then instructed to make these same ratings while focusing only on the target's eye region, in memory. This complemented the composite system's instructions to focus on the eye region when selecting faces from arrays (Fodarella et al., 2017; Skelton et al., 2020, see Supplementary Materials Section 1.4.2).

In summary, we employed three conditions in this experiment (see Design). We predicted that composites produced following holistic recall would be more accurate than those

following a standard, pre-construction CI. Furthermore, we expected best performance when early recall preceded holistic recall.

METHOD

Participants

Construction: Target-unfamiliar participant-witnesses were 30 staff and students (21 female, 9 male) at the University of Lancashire, UK (M_{age} = 26.0, *SD* = 11.0, *Range* = 18 – 43 years), each financially compensated.

Naming: Target-familiar participant-namers were 27 staff and students (15 female, 12 male) at the University of Lancashire, UK (M_{age} = 33.40, *SD* = 16.1, *Range* = 18 – 68 years), each financially compensated.

Likeness Rating: Target-unfamiliar participant-raters were 18 staff and students (9 female, 9 male) at the University of Lancashire, UK (M_{age} = 41.8, *SD* = 16.1, *Range* = 20 – 70 years), each participating voluntarily.

Apparatus and Stimuli

Construction: Materials were the same 10 characters from Experiment 3, prepared to the same standard (see Supplementary Materials Section 1.1). EvoFIT Version 1.6 was used for construction.

Design

Construction and Naming: Based on our predictions, implementing *Early Recall* alone may not facilitate EvoFIT face construction. Therefore, we simplified the intended 2 × 2 design

for Experiment 2 to three conditions, defined by *Interview Type*: CI, where only face-recall was elicited via a Cognitive Interview (coded as 1); H-CI, where holistic recall was added to the CI (coded as 2); and ER-H-CI, a combined approach where early recall preceded the H-CI (coded as 3). Participants-witnesses and -namers were randomly allocated to construct a single composite, or name the set of composites, arising from one of the three between-subjects levels of *Interview Type*.

Likeness Ratings: As before, a within-subjects design was employed for *Interview Type*.

Procedure

Construction: Participant-witnesses first engaged in target encoding (Supplementary Materials Section 1.1). Dependent on condition assignment, a third of participants then independently undertook a written face-recall attempt, 3-4 hours later (Section 1.2). On return to the laboratory (20 – 28 hours post-encoding), participants then engaged in a standard, practitioner-led CI (Supplementary Materials Section 1.3) or a modified H-CI (Section 1.4.2) before proceeding immediately to EvoFIT construction (Section 1.5.3).

Naming and Likeness Rating: These tasks followed the procedure from previous experiments (See Supplementary Materials Sections 2.1 and 3.1, respectively).

RESULTS

Correct Naming

Target photographs were rarely named incorrectly (N = 11/270, M = 4.07%). GEE indicated that the odds of a correct response differed by *Interview Type* (1 = CI, 2 = H-CI, 3 = ER-H-CI) [$\chi_I^2(2) = 80.03$, $p < .001$] (Table 10). Parameter estimates (Table 11) revealed differences

between all three conditions (*p*s < .001), with ER-H-CI performing best, followed by H-CI, and CI performing worst.

[Table 10]

[Table 11]

Mistaken Naming

Mistaken names were infrequent ($N = 28$, $M = 10.8\%$), and notably lower in the ER-H-CI condition (Table 12). GEE indicated different odds of a mistaken response by *Interview Type* [$\chi_I^2(2) = 7.476$, $p = .024$]. Parameter estimates (Table 12, *Note*) revealed a decrease from CI to ER-H-CI ($p = .007$) and from H-CI to ER-H-CI ($p = .016$), while CI and H-CI were equivalent ($p = .85$).

[Table 12]

Composite Likeness Ratings

Participant-raters were generally target-unfamiliar ($M = 12.7\%$ correct). As before, rating-scale-point endorsements were unequal, and so scale recoding was performed (see Supplementary Materials Section 3.1). Ratings increased markedly by condition, from CI to H-CI to ER-H-CI (Table 13).

[Table 13]

GEE revealed that the odds of rated likeness varied by *Interview Type* [$\chi^2(2) = 166.13$, $p < .001$], with all individual conditions emerging different to each other ($p < .001$, Table 13, *Note*): ER-H-CI was best, then H-CI, and lastly CI.

DISCUSSION

Experiment 4 assessed the utility of early and holistic recall techniques for EvoFIT construction. Results for both correct naming and likeness ratings replicated the benefit of early recall when accompanied by holistic recall. Fewer mistaken names were given for composites (indicating more effective composites) following use of both (cf. one or neither) recall techniques.

COMBINED ANALYSES

This section presents a combined analysis across experiments for the two predictors of interest (*Early Recall* and *Interview Type*), providing an overall estimate of their effect sizes. Table 14 displays a summary of means from each experiment. While previous analyses have incorporated conventional sources of variation for items (stimuli) and participant-namers, the current analysis included a third source of variation: the random effect of participant-witnesses, accounting for potential variability introduced by their individual differences.

[Table 14]

The statistical approach remained consistent with previous analyses, incorporating data from Experiments 2 *to* 4 for *Early Recall*, and from Experiments 2 *and* 4 for *Interview Type*. We

again present analyses by-participants here and by-items in Appendix B. All analyses included the random effects of both experiment (coded as 1, 2, etc.) and participant-witnesses (a unique code for participants, 1, 2, etc.). Items were coded uniquely between Experiments 2 and 4, but identically for Experiments 3 and 4, as the same stimuli were used. For GEE, experiment and participant-witnesses were added as between-subject variables in the by-participants analysis, while items were treated as within-subjects (compared to between-subjects in the by-items analysis).

A. Early Recall

(a) For Correct Naming, *Early Recall* [$\chi_1^2(1) = 33.67$, $p < .001$] was retained in the model: Early Recall produced higher correct naming rates than No Early Recall with a medium effect size [$B = 0.84$, $SE(B) = 0.15$, $Exp(B) = 2.32$ (1.74, 3.09)]. This IV was retained in the model by-items.

(b) For Mistaken Naming, *Early Recall* [$\chi_1^2(1) = 3.98$, $p = .046$] was retained: Early Recall produced lower mistaken naming rates than No Early Recall, with a small effect [$B = -0.84$, $SE(B) = 0.15$, $1/Exp(B) = 1.38$ (1.00, 1.90)]. This IV was not retained in the model by-items.

B. Interview Type

(a) For Correct Naming, *Interview Type* [χ_1^2 (1) = 10.93, $p < .001$] was retained: H-CI produced higher correct naming rates than CI, with a small effect size [$B = 0.58$, $SE(B) = 0.18$, $Exp(B) = 1.79$ (1.27, 2.53)]. This IV was not retained in the model by-items.

(b) For Mistaken Naming, *Interview Type* [$\chi_1^2(1) = 3.78$, $p = .052$] was retained: H-CI was associated with marginally higher mistaken naming rates than CI, with a small effect size [*B* = 0.43, *SE(B)* = 0.22, *Exp(B)* = 1.53 (1.00, 2.35)]. This IV was not retained in the model by-items.

GENERAL DISCUSSION

Experiment 1 assessed how retention interval affects PRO-fit construction. Results showed that immediately-constructed composites were most effective, with correct naming and likeness ratings decreasing over time and mistaken naming increasing after 1 week. Of practical importance, PRO-fit composites became largely ineffective at forensically-typical delays, a trend likely to generalise to other feature systems. (e.g., E-FIT, FACES, Frowd et al., 2015).

These findings align with research suggesting that effective composite construction requires sustained access to facial detail (e.g., Brown et al., 2020), which diminishes over time (e.g., Ellis et al., 1980). While Brown et al. (2017) suggest that early practitioner-led verbal elicitation of face-recall (3-4 hours after encoding) can retain access to these details, this implementation depends on practitioner availability. Additionally, verbal recall may be less accurate than written recall (e.g., Miura & Matsuo, 2021), perhaps producing a recoded verbal template that less effectively guides composite construction (Meissner et al., 2001)[6].

[6] Indeed, a follow-up study to Experiment 4 directly assessed this suggestion (Appendix F). Here the effects of written and verbal early recall were compared when both were followed by holistic recall and EvoFIT construction. While the effect of early written recall was replicated for correct naming [*Exp(B)* = 2.68: medium effect = ~2.50], early verbal recall exhibited only a small, non-significant effect [*Exp(B)* = 1.18: small effect = ~1.50].

Our work thus explored whether adding a self-administered written recall attempt 3-4 hours after encoding could improve construction after typical forensic delays.

This technique consistently improved correct naming and likeness ratings across PRO-fit (Experiment 2), Sketch (Experiment 3) and EvoFIT (Experiment 4) systems. For all systems, composite naming and likeness ratings in 'baseline' conditions, which followed standard construction practices, were comparable to those reported in previous work (e.g., Frowd, 2021) and thus variations in these indices can be linked to the implementation of our novel procedures. We particularly expected early recall to benefit PRO-fit composites, where consolidated feature memory might facilitate both initial feature selection and later fine-grained editing [e.g., Frowd et al., 2012a; for correct naming, *Exp(B)* = 2.71: medium effect = ~2.50]. The technique also reduced the odds of a mistaken name being given, emerging as a small, consistent effect in the combined analysis by-participants.

We predicted that the technique might benefit EvoFIT construction to a lesser degree. While consolidated feature-memory might support fine-grained image editing, it does not assist in the initial whole-face selection from arrays (e.g., Frowd et al., 2008, 2012a). Therefore, Experiment 4 did not contain an 'early-recall only' condition. However, when early recall was implemented alongside holistic recall, the benefit for correct naming was larger for EvoFIT than PRO-fit [*Exp(B)* = 2.98] (H-CI compared to ER-H-CI)], while also reducing mistaken identifications compared to when either technique was used alone for PRO-fit.

For artists' sketches, early recall again conferred an advantage, albeit smaller [*Exp(B)* = 1.73: small effect = ~1.50], perhaps because witnesses' preferentially direct artists to sketch groups of features rather than individual features (e.g., Davies & Little, 1990; Laughery et al., 1986;

though *see* Kuivaniemi-Smith, 2023). Further, sketch composites attracted a significantly higher proportion of mistaken names than PRO-fits or EvoFITs. This suggests that following early recall, sketches may accurately represent global feature shapes (as indicated by likeness ratings) but lack fine-grained textural information, leading to activation of related identities during recognition attempts. This higher retrieval of mistaken versus correct names accords with models of *face space* where more generic / less perceptually distinct faces cluster centrally and can be simultaneously activated during recognition attempts, making differentiation between identities difficult (e.g., Burton et al., 1990; Valentine, 1991).

The system-wide benefit of early recall likely does solely reflect early enhancement of face memory. If it did, a face-recall Cognitive Interview soon after encoding should facilitate subsequent construction regardless of later recall attempts. However, Brown et al. (2017) found no such facilitation when participant-witnesses recalled the face only *once*, at a 3–4-hour retention interval, before PRO-fit construction a day later, without a preceding CI. Instead, the first (early) face recall seems to instate a feature-based processing style that enhances output during the second recall immediately before construction. This carry-over represents the testing effect, explained by TAP (e.g., Adesope et al., 2017; Roediger & Karpicke, 2006; Yang et al., 2021). It is unclear whether the strength of this effect may be impacted when the time between initial face encoding and *test* (i.e., initial and subsequent recall attempts) is differently negotiated (e.g., Odinot & Wolters, 2006; Whitten & Bjork, 1977). However, there is some data available on this issue: In a follow-up study (Appendix G), delaying early recall to 24 hours post-encoding resulted in a significant but smaller benefit [*Exp(B)* = 2.23] than when early recall occurred after 3-4 hours (Experiment 4) [*Exp(B)* = 2.98], suggesting a stronger testing effect at shorter retention intervals.

Correct naming also increased when participants reflected on the face's perceived character before construction (holistic recall). The advantage was similar in magnitude for PRO-fit (Experiment 2; [*Exp(B)* = 1.74]) and EvoFIT, when Skelton et al's (2020) eye-region focus interview was adopted (Experiment 4; [*Exp(B)* = 2.02]). Accounting for participant variability, the cross-experiment effect was small and reliable by-participants but not by-items using GEE, though medium [*Exp(B)* = 2.49] when using GLMM (Appendix C).

Following Transfer Appropriate Processing principles, pre-construction holistic recall may specifically prime recognition, rather than recall, mechanisms; the two often considered separable (e.g., Wells & Hryciw, 1984). For feature systems, it may help witnesses to assess when the created image resembles the target (e.g., Frowd et al., 2008, 2012a, 2015). Indeed, correct naming rates for Experiment 2 indicate that early and holistic recall manipulations separately and additively improve PRO-fit effectiveness, although likeness ratings suggested some interaction. For holistic construction, however, holistic recall may play a further necessary role when early-recall protocols are implemented. Here, early-recall may entrench a feature-processing style that facilitates late construction activities (i.e., making fine-grained feature edits to enhance likeness; Frowd et al., 2012a), but impedes earlier ones, specifically, selection of faces from arrays that best resemble the target (e.g., Brown & Lloyd-Jones, 2002; Frowd & Fields, 2011; Maclin, 2002). The benefits of early-recall for holistic construction may then only be observed when holistic recall occurs between feature recall and construction. Positioned here it may temporarily recalibrate witnesses to a more appropriate processing style (e.g., Schooler & Engstler-Schooler, 1990), without compromising the existence of, or later access to, recall-consolidated feature memory (Fodarella et al., 2021; Skelton et al., 2020). Supporting this proposal, results showed higher correct naming and

likeness ratings, and lower mistaken naming, for EvoFITs constructed using this combined approach (cf. holistic recall, only; see also Appendix F).

Strengths, Limitations and Future Work

Sample characteristics varied considerably across experiments, according to age range, gender split, locality, and sample composition (i.e., staff-to-student ratios). While this may be viewed as a limitation, previous similar work suggests that individual differences within participant-witness and participant-namer samples typically have little impact on key experimental outcomes (e.g., Frowd et al., 2015), particularly when appropriate condition randomisation has been employed (*see* Supplementary Materials). Further, significant fixed effects for correct naming continued to be returned in a combined by-participants analysis, when variability across participant-witnesses was controlled (i.e., by adding participant-witnesses as a random effect to the model [*Exp(B)* = 2.32]). Our combined analysis also found significant fixed-effects by-items (Appendix B), despite cross-experimental differences in the target pools from which our identities were drawn, and the specific identities used. This suggests that our findings will generalise to other stimuli (i.e., real-world identities).

A potentially more relevant limitation was the lack of participant supervision between target encoding and face construction. While some participants may not have thought about the face during this period, awareness of the upcoming task may have encouraged rehearsal. In particular, those who completed early recall might have intuited that this retrieval attempt was designed to improve their performance, and so they may have reviewed and / or replicated their descriptions before construction. This potential behavioural variability complicates conclusions about the utility of a *single*, *specifically-timed* self-administered

interview. Future researchers should explicitly track how often participants intentionally (or spontaneously, e.g., Turtle & Yuille, 1994) thought about the face, reviewed and / or replicated their descriptions during the retention interval, and then include this variable as a moderator in analyses.

We also analysed responses from participant-namers and -raters using Generalized Linear Mixed Models (see Appendices C-E). Despite their decades-long availability (e.g., Agresti et al., 2000), GLMMs' complexity has limited their adoption (Bolker et al., 2009). However, when properly applied, they offer substantial advantages over ANOVA and GEE by simultaneously considering by-participants and by-items factors, avoiding difficulties in reconciling disparate trends. Like GEE, statistical design remains crucial—particularly the ability to detect forensically-useful medium effect sizes for naming with good power. We evaluate this statistical approach and compare GLMM to GEE in Appendix D. We conclude that GLMMs' single inferential outcome provides greater parsimony while elegantly accounting for numerous sources of variance.

Conclusions

This research demonstrates substantial benefits of a novel technique designed to preserve and consolidate feature memory across forensically-typical delays. The method—having witnesses provide written recall before verbal recall and construction a day later—is simple to implement without practitioner oversight. It appears effective across different construction systems (PRO-fit, Sketch and EvoFIT). Given that these systems are representative of those used in forensic practice, we would expect our results to generalise to other feature and holistic interfaces. Indeed, this proposal is currently being trialled by six police forces, whereby investigating officers are requesting witnesses and victims to write a detailed

description of the offender's face, with composite construction arranged later with a practitioner. Moreover, combining early recall with holistic techniques shows additional benefits for both feature and holistic composites.

Declarations:

Funding Details: The authors report no receipt of external funding for this work.

Disclosure Statement: The authors report there are no competing interests to declare.

Data Availability Statement: Raw data for these experiments were generated at the Universities of Lancashire, Leeds and Dundee. Derived data supporting the findings of this study are available from the corresponding author [EP] on request.

Acknowledgement: We would like to thank Karen Grace-Martin from The Analysis Factor (www.theanalysisfactor.com) for her valuable comments on a previous version of this paper.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701.

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-effects modelling of categorical response data. *Sociological Methodology, 30*(1), 27–80.

Ashkenazi, T., & Fisher, R. P. (2022). Field test of the cognitive interview to enhance eyewitness and victim memory, in intelligence investigations of terrorist attacks. *Journal of Applied Research in Memory and Cognition, 11*, 200–208.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Setevens, M. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24*, 127–135.

Brown, C., & Lloyd-Jones, T. J. (2002). Verbal overshadowing in a multiple face presentation paradigm: Effects of description instruction. *Applied Cognitive Psychology, 16*(8), 873–885.

Brown, C., Portch, E., & Frowd, C. D. (2017). Tell me again about the face: Using repeated interviewing techniques to improve feature-based facial composite technologies. In *Proceedings of the 2017 Seventh International Conference on Emerging Security Technologies (EST)* (pp. 38–43). Institute of Electrical and Electronics Engineers.

Brown, C., Portch, E., Nelson, L., & Frowd, C. D. (2020). Reevaluating the role of verbalization of faces for composite production: Descriptions of offenders matter! *Journal of Experimental Psychology: Applied, 26*, 248–265.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*, 361–380.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Damin, E. A. (2018). The impact of early recall on the efficiency of face composite construction using the EvoFIT system [Master's thesis]. University of Lancashire.

Davies, G. M., & Little, M. (1990). Drawing on memory: Exploring the expertise of a police artist. *Medical Science and the Law, 30*(4), 345–354.

Ebbesen, E. B., & Rienick, C. B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology, 5*, 745–762.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. *Journal of Police Science and Administration, 8*, 101–106.

Erickson, W. B., Brown, C., Portch E., Lampinen, J. M., Marsh, J. M., Fodarella, C., Petkovic, A., Coultas, C., Newby, A., Data, L., Hancock, P. J. B., & Frowd, C. D. (2022). The impact of weapons and unusual objects on the construction of facial composites. *Psychology, Crime & Law*, *30*(3), 207–228.

Field, A. (2018). *Discovering statistics using SPSS* (5th ed.). Sage.

Fisher, R. P., Geiselman, R. E., Raymond, D. S., Jurkevich, L. M., & Warhaftig, M. L. (1987). Enhancing enhanced eyewitness memory: Refining the cognitive interview. *Journal of Police Science and Administration, 15*, 291–297.

Fodarella, C., Frowd, C. D., Warwick, K., Hepton, G., Stone, K., Date, L., & Heard, P. (2017). Adjusting the focus of attention: Helping witnesses to evolve a more identifiable composite. *Forensic Research & Criminology International, 5*(1), Article 00143.

Fodarella, C., Marsh, J. E., Chu, S., Athwal-Kooner, P., Jones, H. S., Skelton, F. C., Wood, E., Jackson, E., & Frowd, C. D. (2021). The importance of detailed context reinstatement for the production of identifiable composite faces from memory. *Visual Cognition, 29*(3), 180–200.

Frowd, C. D. (2021). Forensic facial composites. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in forensic facial-recognition* (pp. 34–64). Taylor and Francis.

Frowd, C. D. (2023). Eyewitnesses and the use and application of cognitive theory. In G. Davey (Ed.), *Introduction to Applied Psychology* (pp. 207–232). BPS Wiley-Blackwell.

Frowd, C. D., Bruce, V., Ness, H., Bowie, L., Thomson-Bogner, C., Paterson, J., McIntyre, A., & Hancock, P. J. B. (2007). Parallel approaches to composite production. *Ergonomics, 50*, 562–585.

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

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

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

Frowd, C. D., Erickson, W. B., Lampinen, J. L., Skelton, F. C., McIntyre, A. H., & Hancock, P. J. B. (2015). A decade of evolving composite techniques: Regression- and meta-analysis. *Journal of Forensic Practice, 17*, 319–334.

Frowd, C. D., & Fields, S. (2011). Verbalisation effects in facial composite production. *Psychology, Crime & Law, 17*, 731–744.

Frowd, C. D., McQuiston-Surrett, D., Kirkland, I., & Hancock, P. J. B. (2005c). The process of facial composite production. In A. Czerederecka, T. Jaskiewicz-Obydzinska, R. Roesch & J. Wojcikiewicz (Eds.), *Forensic psychology and law* (pp. 140–152). Institute of Forensic Research Publishers.

Frowd, C. D., Nelson, L., Skelton, F. C., Noyce, R., Atkins, R., Heard, P., Morgan, D., Fields, S., Henry, J., McIntyre, A., & Hancock, P. J. B. (2012a). Interviewing techniques for Darwinian facial composite systems. *Applied Cognitive Psychology, 26*, 576–584.

Frowd, C. D., Pitchford, M., Bruce, V., Jackson, S., Hepton, G., Greenall, M., McIntyre, A., & Hancock, P. J. B. (2010). The psychology of face construction: Giving evolution a helping hand. *Applied Cognitive Psychology, 25*, 195–203.

Frowd, C. D., Pitchford, M., Skelton, F. C., Petkovic, A., Prosser, C., & Coates, B. (2012b). Catching even more offenders with EvoFIT facial composites. In *Proceedings of the 2012 Third International Conference on Emerging Security Technologies (EST)* (pp. 20–26). Institute of Electrical and Electronics Engineers.

Frowd, C. D., Portch, E., Killeen, A., Mullen, L., Martin, A. J., & Hancock, P. J. B. (2019). EvoFIT facial composite images: A detailed assessment of impact on forensic practitioners, police investigators, victims, witnesses, offenders and the media. In *Proceedings of the 2019 Eighth International Conference on Emerging Security Technologies (EST)* (pp. 1–7). Institute of Electrical and Electronics Engineers.

Frowd, C. D., Skelton, F. C., Atherton, C., Pitchford, M., Hepton, G., Holden, L., McIntyre, A., & Hancock, P. J. B. (2012c). Recovering faces from memory: The

distracting influence of external facial features. *Journal of Experimental Psychology: Applied, 18*, 224–238.

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

Gabbert, F., Hope, L., & Fisher, R. (2009). Protecting eyewitness evidence: Examining the efficacy of a self-administered interview. *Law and Human Behavior, 33*, 298–307.

Gill, J., & King, G. (2004). What to do when your hessian is not invertible. *Sociological Methods & Research, 33*, 54–87.

IBM (2020). Can one get one-tailed tests in Logistic Regression by dividing significance levels in half? IBM Support, Document 422407. https://www.ibm.com/support/pages/can-one-get-one-tailed-tests-logistic-regression-dividing-significance-levels-half

IBM (2021). *IBM SPSS Statistics for Windows* (Version 29.0) [Computer software]. IBM Corp.

Kellogg, R. T. (2007). Are written and spoken recall of text equivalent? *The American Journal of Psychology*, *120*(3), 415–428.

Kuivaniemi-Smith, H. J. (2023). *Understanding and improving the effectiveness of sketch facial composites* [Doctoral dissertation]. University of Lancashire.

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

MacLin, M. K. (2002). The effects of exemplar and prototype descriptors on verbal overshadowing. *Applied Cognitive Psychology, 16*(8), 929–936.

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition, 29*, 176–186.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, Article 104092.

Miura, H., & Matsuo, K. (2021). Does writing enhance recall and memory consolidation? Revealing the factor of effectiveness of the self-administered interview. *Applied Cognitive Psychology, 35*(5), 1338–1343.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Odinot, G., Memon, A., La Rooy, D., & Millen, A. (2013). Are two interviews better than one? Eyewitness memory across repeated cognitive interviews. *PLOS ONE, 8*, Article e76305.

Odinot, G., & Wolters, G. (2006). Repeated recall, retention interval and the accuracy-confidence relation in eyewitness memory. *Applied Cognitive Psychology, 20*, 973–985.

Osborne, J. W. (2016). *Regression & linear modeling: Best practices and modern methods*. Sage Publications.

Pitchford, M., Green, D., & Frowd, C. D. (2017). The impact of misleading information on the identifiability of feature-based facial composites. In *Proceedings of the 2017 Seventh International Conference on Emerging Security Technologies (EST)* (pp. 185–190). Institute of Electrical and Electronics Engineers.

Reed, P., & Wu, Y. (2013). Logistic regression for risk factor modelling in stuttering research. *Journal of Fluency Disorders, 38*, 88–101.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Sauerland, M., Krix, A. C., van Kan, N., Glunz, S., & Sak, A. (2014). Speaking is silver, writing is golden? The role of cognitive and social factors in written versus spoken witness accounts. *Memory & Cognition, 42*, 978–992.

Sauerland, M., & Sporer, S. L. (2011). Written vs. spoken eyewitness accounts: Does modality of testing matter? *Behavioral Sciences and the Law, 29*, 846–857.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36–71.

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

Skelton, F. C., Frowd, C. D., Hancock, P. J. B., Jones, H. S., Jones, B. C., Fodarella, C., Battersby, K., & Logan, K. (2020). Constructing identifiable composite faces: The importance of cognitive alignment of interview and construction procedure. *Journal of Experimental Psychology: Applied, 26*, 507–521.

Solomon, C. J., & Gibson, S. J. (2013). Developments in forensic facial composites. In X. Mallett & T. Blythe (Eds.), *Advances in forensic human identification* (pp. 235–270). CRC Press.

Sporer, S. L., & Martschuk, N. (2014). The reliability of eyewitness identifications by the elderly: An evidence-based review. In M. P. Toglia, D. F. Ross, J. Pozzulo, & E. Pica (Eds.), *The elderly eyewitness in court* (pp. 3–37). Psychology Press.

Tredoux, C.G., Rosenthal, Y., da Costa, L. & Nunez, D. (1999). Evaluation of an eigenface-based composite system. Paper presented at 3rd meeting of the Society for Applied Research in Memory and Cognition. Boulder, Colorado, 10 July 1999.

Turtle, J. W., & Yuille, J. C. (1994). Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hypermnesia. *Journal of Applied Psychology, 79*(2), 260–271.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*(2), 161–204.

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

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16*(4), 465–478.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435.

Appendix A: Additional information for Analyses using Generalized Estimating Equations

To keep the Results' sections uncluttered, information about the analyses by-participants for each final model are presented here. See Appendix B for associated analyses conducted by-items (i.e., for the identities of the stimuli), and Appendix E for a table of statistical comparisons.

Experiment 1

Naming. Data were analysed for composites for which participant-namers correctly named the corresponding target photographs (N = 390 out of 400).

(a) Correct. Information Criteria (*QIC* = 276.6, *QICC* = 280.0) and Intercept [B = -0.78, *SE(B)* = 0.17].

(b) Mistaken. Information Criteria (*QIC* = 504.3, *QICC* = 498.8) and Intercept [B = -0.87, *SE(B)* = 0.29].

Likeness. Data were analysed for composites for which participant-raters did not correctly name the corresponding target photographs (N = 635 out of 680). Note that the mean is used here (and elsewhere in the paper) for likeness ratings as this measure of central tendency clearly expresses group differences. Unadjusted means (i.e., scale range 1 – 15, without recoding): Immediate = 5.2, 3-4 hours = 4.5, 2 days = 4.3 and 1 week = 3.8. Threshold rating (scale) values of B were: *1* = -1.51, *2* = -0.87, *3* = -0.38, *4* = 0.03, *5* = 0.40, *6* = 0.72 and *7* = 1.13.

Experiment 2

Naming. Data were analysed for composites for which participant-namers correctly named the corresponding target photographs (N = 400 out of 400).

(a) Correct. Information Criteria (*QIC* = 496.8, *QICC* = 500.9) and Intercept [*B* = -1.40, *SE(B)* = 0.12].

(b) Mistaken. Information Criteria (*QIC* = 400.1, *QICC* = 398.0) and Intercept [*B* = -1.74, *SE(B)* = 0.24].

Likeness. Data were analysed for composites for which participant-raters did not correctly name the corresponding target photographs (*N* = 708 out of 720). Unadjusted means (without recoding): CI / No Early Recall = 2.2, CI / Early Recall = 2.5, H-CI / No Early Recall = 2.1, H-CI / Early Recall = 3.6. Threshold rating values of *B* (*1* = -0.48, *2* = 0.91, *3* = 1.75, *4* = 2.69).

Experiment 3

Naming. Data were analysed for composites for which participant-namers correctly named the corresponding target photographs (*N* = 241 out of 250).

(a) Correct. Information Criteria (*QIC* = 326.9, *QICC* = 326.8) and Intercept [*B* = -0.64, *SE(B)* = 0.20].

(b) Mistaken. Information Criteria (*QIC* = 335.7, *QICC* = 335.9) and Intercept [*B* = -0.09, *SE(B)* = 0.18].

Likeness. Data were analysed for composites for which participant-raters did not correctly name the corresponding target photographs (*N* = 358 out of 360). Unadjusted means (i.e., without recoding): No Early Recall = 3.2, Early Recall = 3.9. Threshold rating values of *B* (*1* = -1.67, *2* = -0.62, *3* = 0.09, *4* = 0.93, *5* = 2.18).

Experiment 4

Naming. Data were analysed for composites for which participant-namers correctly named the corresponding target photographs (*N* = 259 out of 270).

(a) Correct. Information Criteria (*QIC* = 327.9, *QICC* = 332.1) and Intercept [*B* = -0. 89, *SE(B)* = 0.15].

(b) Mistaken. Information Criteria (*QIC* = 173.6, *QICC* = 174.3) and Intercept [*B* = -1.68, *SE(B)* = 0.28].

Likeness. Data were responses to composites for which participant-raters did not correctly name the corresponding target photographs (*N* = 471 out of 540). Unadjusted means: CI = 2.7, H-CI = 3.6, ER-H-CI = 5.1. Intercept [*B* = -0.99, *SE(B)* = 0.39], Information Criteria (*QIC* = 348.3, *QICC* = 332.3) and Threshold rating values of *B* (*3* = 1.33, *4* = 2.77).

Combined Analyses

A. Early Recall

Naming. Data were analysed for composites for which participant-namers correctly named the corresponding target photographs (*N* = 817 out of 830).

(a) Correct. Intercept [*B* = -0.75, *SE(B)* = 0.11] and Information Criteria (*QIC* = 1083.5, *QICC* = 1083.5).

(b) Mistaken. Intercept [*B* = -0.94, *SE(B)* = 0.11] and Information Criteria (*QIC* = 920.5, *QICC* = 920.5).

B. Interview Type

Naming. Data were analysed for composites for which participant-namers correctly named the corresponding target photographs (*N* = 571 out of 580).

(a) Correct. Intercept [*B* = -0.86, *SE(B)* = 0.13] and Information Criteria (*QIC* = 740.9, *QICC* = 740.9).

(b) Mistaken. Intercept [*B* = -1.72, *SE(B)* = 0.17] and Information Criteria (*QIC* = 545.2, *QICC* = 545.2).

Appendix B: By-items Generalized Estimating Equations

The GEE analyses presented in the paper follow an established approach for analysing responses to composites (e.g., Frowd et al., 2013). They assess the effectiveness of constructed composites (via naming and rated likeness tasks) with respect to participants for the various IVs under investigation. The approach thus indicates the extent to which results generalise to other participants. However, to avoid the risk of making a stimuli-as-a-fixed-effects fallacy (Clark, 1973), here we conducted analyses that focused on the individual items of stimuli, to give a measure of how results generalise to other identities. Thus, analyses by-items were conducted in the same way as by-participants. This included an IV or interaction being maintained in the model if $\alpha < .1$. We also conducted combined analyses, as before, that included a third important source of random variation: the effect of participant-witnesses. In the following, due to space constraints, results are again presented concisely (including without use of tables); details of the omnibus test are stated first, followed by relevant post-hoc test(s) and simple-main effects.

For the individual experiments, the analyses by-items presented here led to the same pattern of significant and non-significant differences as by-participants analyses except that, in Experiment 2, mistaken naming was marginally lower by-items in the omnibus test ($p = .066$) following early (cf. no early) recall, while this difference was not significant by-participants ($p = .15$). Also, in the Combined Analyses, results for early (cf. no early) recall were consistent for correct and mistaken naming, but there were inconsistencies for H-CI (cf. CI), which were significant by-participants but not by-items, presumably as this predictor emerged with an effect size that was smaller than the planned medium effect for the analysis.

The authors note that an alternative solution to the potential issue arising from conducting separate by-participant and by-item analyses are presented in Appendix C, where such inconsistencies are avoided by using GLMM.

Experiment 1

(a) Correct Naming

Retention Interval was retained in the model [$\chi_2^2(3) = 20.35$, $p < .001$]. Conducting Reverse Helmert contrasts revealed that the odds of a correct response was lower after (i) 3-4 hours than immediate [*SE(M)* = 0.06, $p = .008$], (ii) 2 days than the shorter (immediate and 3-4 hour) delays [*SE(M)* = 0.04, $p = .007$] and (iii) 1 week than all other delays combined [*SE(M)* = 0.03, $p < .001$]. Details of this model were Intercept [*B* = -0.99, *SE(B)* = 0.27] and Information Criteria (*QIC* = 294.9, *QICC* = 280.0).

(b) Mistaken Naming

Retention Interval was retained [$\chi_2^2(3) = 18.44$, $p < .001$], and Reverse Helmert contrasts revealed that the odds of a mistaken response were higher at 1 week relative to (combined) shorter delays [*SE(M)* = 0.06. $p < .001$]; other contrasts were *ns* (*p*s > .28). Intercept [*B* = -0.89, *SE(B)* = 0.24] and Information Criteria (*QIC* = 500.6, *QICC* = 498.8).

(c) Likeness Ratings

Retention Interval was again retained in the model [$\chi_2^2(3) = 10.13$, $p = .018$]. In three separate models[2], lower odds of rated likeness was found for composites constructed after (i) 3-4 hours (marginally) compared to immediate [*B* = -0.32, *SE(B)* = 0.18, $p = .079$] and (ii) 1 week than all other delays [*B* = -0.40, *SE(B)* = 0.16, $p = .011$]; no difference in odds were found between 2 days and the shorter (immediate and 3-4 hour) delays [*B* = -0.16, *SE(B)* =

0.16, p = .32]. Threshold rating values of *B* (*1* = -1.53, *2* = -0.87, *3* = -0.38, *4* = 0.02, *5* = 0.39, *6* = 0.70, *7* = 1.11).

Experiment 2

(a) Correct Naming

In a full-factorial model, the interaction was removed [$\chi_2^2(1) = 0.01$, p = .916, *1/Exp(B)* = 1.03]. The resulting, final model comprised *Early Recall* [$\chi_2^2(1) = 28.71$, $p < .001$], with Early Recall > No Early Recall [*B* = 1.00, *SE(B)* = 0.19, $p < .001$, *Exp(B)* = 2.71† (1.88, 3.90)]; and *Interview Type* [$\chi_2^2(1) = 9.69$, $p < .001$], with H-CI > CI [*B* = 0.56, *SE(B)* = 0.18, p = .003, *Exp(B)* = 1.74‡ (1.23, 2.47)]. For this final model, Intercept [*B* = -1.40, *SE(B)* = 0.47] and Information Criteria (*QIC* = 534.7, *QICC* = 500.9).

† For the avoidance of doubt, this effect size (to two d.p.) by-items (2.708) is exactly the same as that found by-participants (2.706).

‡ For the avoidance of doubt, this effect size (to three d.p.) by-items (1.741) is exactly the same as that found by-participants (1.741).

(b) Mistaken Naming

The interaction was removed from the full-factorial model [$\chi_2^2(1) = 1.40$, p = .238, 1/*Exp(B)* = 1.82]**Error! Bookmark not defined.**. In the resulting, final model, both predictors were retained: *Early Recall* [$\chi_2^2(1) = 3.40$, p = .066], with No Early Recall marginally > Early Recall [*B* = 0.45, *SE(B)* = 0.24, *Exp(B)* = 1.56 (0.97, 2.51)]; and *Interview Type* [$\chi_2^2(1) = 6.66$, p = .010], with H-CI > CI [*B* = 0.64, *SE(B)* = 0.25, *Exp(B)* = 1.89 (1.17, 3.08)]. Intercept [*B* = -1.52, *SE(B)* = 0.35] and Information Criteria (*QIC* = 409.5, *QICC* = 396.9).

(c) Likeness Ratings

In a full-factorial model, the interaction was retained [$\chi_2^2(1) = 27.76$, $p < .001$]; IVs were *Interview Type* [$\chi_2^2(1) = 16.65$, $p < .001$] and *Early Recall* [$\chi_2^2(1) = 67.55$, $p < .001$]. For the interaction: (i) Early Recall > No Early Recall at each level of interview (ps ≤ .007) but (ii) H-CI > CI with Early Recall ($p < .001$) but not when Early Recall was omitted ($p = .33$). In detail: Early Recall > No Early Recall: CI [$B = 0.42$, $SE(B) = 0.16$, $p = .007$, $Exp(B) = 1.53$ (1.12, 2.08)] and H-CI [$B = 1.66$, $SE(B) = 0.19$, $p < .001$, $Exp(B) = 5.26$ (3.64, 7.58)]. H-CI > CI: Early Recall [$B = 1.09$, $SE(B) = 0.18$, $p < .001$, $Exp(B) = 2.97$ (2.11, 4.18)] and No Early Recall (ns) [$B = -0.15$, $SE(B) = 0.15$, $p = 0.33$, $1/Exp(B) = 1.16$ (0.85, 1.57)]. For this model, Threshold rating values of B ($1 = -0.50$, $2 = 0.73$, $3 = 1.40$, $4 = 2.24$).

Experiment 3

(a) Correct Naming

Early Recall [$\chi_2^2(1) = 5.77$, $p = .016$]: Early Recall > No Early Recall [$B = 0.57$, $SE(B) = 0.24$, $\chi_2^2(1) = 5.77$, $p = .016$, $Exp(B) = 1.76$ (1.11, 2.80)]. Intercept [$B = -0.68$, $SE(B) = 0.36$] and Information Criteria ($QIC = 339.5$, $QICC = 326.8$).

(b) Mistaken Naming

Early Recall [$\chi_2^2(1) = 0.55$, $p = .458$]: Thus, Early Recall was equivalent to No Early Recall [$B = -0.17$, $SE(B) = 0.23$, $X_2^2(1) = 0.55$, $1/Exp(B) = 1.19$ (0.75, 1.88)]. Intercept [$B = -0.06$, $SE(B) = 0.32$] and Information Criteria ($QIC = 348.7$, $QICC = 335.9$).

(c) Likeness Ratings

Early Recall [$\chi_2^2(1) = 14.32$, $p < .001$]: Early Recall > No Early Recall [$B = 0.70$, $SE(B) = 0.19$, $p < .001$, $Exp(B) = 2.02$ (1.40, 2.91)]. Threshold rating values of B ($1 = -1.54$, $2 = -0.56$, $3 = 0.15$, $4 = 0.97$, $5 = 2.12$).

Experiment 4

(a) Correct Naming

Interview Type [$\chi_2^2(2) = 38.90$, $p < .001$]: H-CI > CI [$B = 0.78$, $SE(B) = 0.29$, $p = .008$, $Exp(B) = 2.17$ (1.22, 3.87)], ER-H-CI > CI [$B = 1.88$, $SE(B) = 0.30$, $p < .001$, $Exp(B) = 6.55$ (3.61, 11.90)] and ER-H-CI > H-CI [$B = 1.10$, $SE(B) = 0.29$, $p < .001$, $Exp(B) = 3.01$ (1.72, 5.27)]. Intercept [$B = -0.99$, $SE(B) = 0.39$] and Information Criteria ($QIC = 348.3$, $QICC = 332.3$).

(b) Mistaken Naming

Interview Type [$\chi_2^2(2) = 6.52$, $p = .038$]: CI > H-CI (ns) [$B = 0.16$, $SE(B) = 0.43$, $p = .71$, $Exp(B) = 1.17$ (0.50, 2.75)], CI > ER-H-CI [$B = 1.66$, $SE(B) = 0.66$, $p = .012$, $Exp(B) = 5.26$ (1.44, 19.19)] and H-CI > ER-H-CI [$B = 1.50$, $SE(B) = 0.66$, $p = .024$, $Exp(B) = 4.48$ (1.22, 16.47)]. Intercept [$B = -1.69$, $SE(B) = 0.30$] and Information Criteria ($QIC = 174.1$, $QICC = 174.3$).

(c) Likeness Ratings

Interview Type [$\chi_2^2(2) = 167.24$, $p < .001$]: H-CI > CI [$B = 1.48$, $SE(B) = 0.25$, $p < .001$, $Exp(B) = 4.37$ (2.69, 7.11)], ER-H-CI > CI [$B = 3.61$, $SE(B) = 0.28$, $p < .001$, $Exp(B) = 36.82$ (21.16, 64.06)] and ER-H-CI > H-CI [$B = 2.13$, $SE(B) = 0.24$, $p < .001$, $Exp(B) = 8.40$ (5.26, 13.51)]. Threshold rating values of B ($3 = 1.33$, $4 = 2.78$).

Combined Analyses

A. Early Recall

(a) Correct Naming. *Early Recall* was retained [$\chi_2^2(1) = 7.49$, $p = .006$], with Early Recall higher than No Early Recall with a medium effect [$B = 0.86$, $SE(B) = 0.31$, $Exp(B) = 2.36$ (1.28, 4.37)]. Intercept [$B = -0.75$, $SE(B) = 0.23$] and Information Criteria ($QIC = 1098.7$, $QICC = 1083.5$). The IV was retained in the model by-participants.

(b) Mistaken Naming. *Early Recall* was removed from the model [$\chi_2^2(1) = 1.22$, $p = .270$]. Thus, Early Recall was equivalent to No Early Recall [$B = -0.37$, $SE(B) = 0.34$, $1/Exp(B) = 1.45$ (0.75, 2.82)]. Intercept [$B = -0.98$, $SE(B) = 0.23$] and Information Criteria ($QIC = 933.5$, $QICC = 921.3$). This IV was retained in the model by-participants.

B. Interview Type

(a) Correct naming. *Interview Type* was greater than alpha [$\chi_2^2(1) = 2.30$, $p = .129$] and so was removed: H-CI was equivalent to CI [$B = 0.60$, $SE(B) = 0.39$, $Exp(B) = 1.82$ (0.84, 3.94)]. Intercept [$B = -1.72$, $SE(B) = 0.34$] and Information Criteria ($QIC = 549.7$, $QICC = 545.2$). This IV was retained in the model by-participants.

(b) Mistaken Naming. *Interview Type* was removed from the model [$\chi_2^2(1) = 1.63$, $p = .202$]: H-CI was equivalent to CI [$B = 0.41$, $SE(B) = 0.32$, $Exp(B) = 1.51$ (0.80, 2.85)]. Intercept [$B = -0.99$, $SE(B) = 0.39$] and Information Criteria ($QIC = 348.3$, $QICC = 332.3$). This IV was retained in the model by-participants.

Appendix C: Generalized Linear Mixed Effects Models (GLMM)

Our approach followed the established statistical method for analysing responses to composites using GEE (e.g., Brown et al., 2017; Frowd et al., 2013; Pitchford et al., 2017). However, we took this opportunity to conduct the analyses using a similar approach, GLMM. This method involves a unified model, one that essentially combines analyses by-participants and by-items. GLMM are considered best practice for hypothesis testing (Barr et al., 2013). As elsewhere, results are presented concisely[7] (with the results of the first experiment presented in more detail).

We followed the statistical method described in Erickson et al. (2022) for GLMM (GENLINMIXED, SPSS Version 29, IBM 2020, 2021). The approach is same as that described in the current paper for GEE with respect to scoring, coding, selection of cases and approach. The main difference between GEE and GLMM is the way in which random effects are handled. GEE models responses as being equally correlated (using an Exchangeable Working Correlation matrix), averaging over *items* in the by-participants analysis, and *participants* in the by-items analysis: in contrast, GLMM de-correlates responses by including random factors for participants and items. More specifically, based on available variance in the data, GLMM fits a random intercept for each participant and for each item, as well as a random slope for any within-subjects predictors that are included in the model.

There are two points to note. Firstly, models were 'maximal'. That is, they included as many random intercepts and random slopes as indicated by the design. They were then simplified, where random effects were only retained in the model for which sufficient variance (σ^2) was

[7] We acknowledge that the results would have been better if again presented as a series of tables; however, this format was not possible due to publication constraints.

available in the data. This approach is best practice (Barr et al., 2013). Note that is not a problem in itself that a random effect cannot be estimated; for example, participant-namers are often sufficiently consistent in their responses that random intercepts for this source of error are not required (cf. items). Overall, this process, when transforming to the response scale (which we do here), leads to inference on the subject with zero random effect. Secondly, due to de-correlation by inclusion of maximal random effects, the covariance type was specified with responses as being independent (achieved in SPSS by selecting *Variance Components*).

Since Robust produced either the same or higher SE values, the same as for GEE, we again selected a Model-based (cf. Robust) setting for the covariance matrix throughout. As described, an independent (cf. exchangeable) covariance structure was selected since the correlated nature of responses had already been taken into account (i.e., in the random-effects model). For SPSS GLMM (vs. GEE), we note that (i) GLMM provides an overall fit of the model (called a "Corrected Model") (cf. GEE), details of which are included in models containing more than one predictor[8], (ii) F replaces X^2 and (iii) AIC and BIC replace QIC and QICC measures for Information Criteria for naming, but neither measure is available for analysing multinomial responses (e.g., from ratings of likeness).

Based on a comparable design to the current experiments (Erickson et al., 2022), our expectation was that inferential analyses would be similar between GEE and GLMM, if not the same—although we note that Random-intercepts-only GLMM (used for naming analyses here) generalize somewhat worse than separate by-participants and by-items tests using GEE (Gill & King, 2004). Our expectation turned out to be true for correct naming; it was also true

[8] For brevity, details of the Corrected Model are omitted for analyses that contain a single predictor since (as is usual with regression analyses) these details are identical to those of the predictor itself.

for mistaken naming, although there was an issue with model validity in Experiment 4 (see below for ways to reduce this issue, such as by using GEE or by increasing sample size). In fact, GLMM were conducted on the simulated correct naming data sets described in Supplementary Materials Section 4.1.1. The outcome was very similar: the predictor of interest remained in the model ($p < .001$ to .025, *Exp(B)* = 2.57); *SE(B)* for the predictor varied from 0.31 to 0.43. Again, all samples were maintained for $\alpha = .1$, and $1\text{-}\beta > 95\%$.

For the supplementary measure, ratings of likeness, analyses involved adjusted (scale-collapsed) data, as described in the paper. The outcome of the inferential analyses was basically the same when random effects were minimal (i.e., when including random intercepts only), but most effects did not emerge (as the relevant predictor was removed from the model) when random effects were maximal (i.e., when also including random slopes). This effect was observed by Erickson et al. (2022). Adding random slopes provides a more accurate model, but this outcome suggests that a larger sample size is necessary to accommodate higher emerging SE when analysing ordinal-level responses. Indeed, the anticipated advantage of increased sample size is illustrated below in the combined analyses.

Experiment 1

(a) Correct Naming.

The (GLMM) model contained *Retention Interval* [$F(3,386) = 7.52$, $p < .001$] as a fixed effect (IV); there was sufficient variability to include random intercepts for items ($\sigma^2 = 0.64$, *SE* = 0.46) but not (due to consistent responses between participant-namers) random intercepts for participant-namers ($\sigma^2 = 0.0$). Other details for this model were Overall Correct Classification (86.9%), Intercept [*B* = -1.08, *SE(B)* = 0.35] and Information Criteria (*AICC* = 2108.7, *BIC* = 2112.7).

Unlike GEE, Reverse Helmert contrasts for GLMM are not available in SPSS and so these post-hoc tests were specified using a dummy-coded variable in three separate models for Retention Interval. Using this approach, the odds of a correct response to composites was worse after (i) 3-4 hours than the previous (immediate) delay [p = .016, *SE(B)* = 0.39, 1/*Exp(B)* = 2.60]**Error! Bookmark not defined.**, (ii) 2 days than the previous (immediate and 3-4 hour) delays [p = .016, *SE(B)* = 0.40, 1/*Exp(B)* = 2.70] and (iii) 1 week than all previous delays [p = .003, *SE(B)* = 0.61, 1/*Exp(B)* = 6.41].

(b) Mistaken Naming

The model retained *Retention Interval* [$F(3,386) = 6.11$, $p < .001$]; the same as for correct naming, it contained random intercepts for items ($\sigma^2 = 0.10$, *SE* = 0.11) only; other details were Overall Correct Classification (67.2%), Intercept [*B* = -0.90, *SE(B)* = 0.25] and Information Criteria (*AICC* = 1711.2, *BIC* = 1715.1).

Reverse Helmert contrasts indicated that the odds of a mistaken response was higher for composites constructed at 1 week relative to (combined) previous delays (*SE* = 0.35. $p <$ .003, *Exp(B)* = 2.81); other contrasts were *ns* (ps > .39).

(c) Likeness Ratings

The model contained *Retention Interval* [$F(3,625) = 5.46$, p = .001]; see below for details of random effects. As before, simulating Reverse Helmet contrasts[2], the odds of rated likeness after (i) 3-4 hours was lower than the previous (immediate) delay [*B* = -0.43, *SE(B)* = 0.20, p = .035], (ii) 2 days was equivalent to the previous (immediate and 3-4 hour) delays [*B* = -0.25, *SE(B)* = 0.18, p = .15], and (iii) 1 week was lower than all previous delays [*B* = -0.53,

SE(B) = 0.17, *p* = .002]. Other details for this model were Threshold rating values of *B* (*1* = -1.84, *2* = -1.04, *3* = -0.42, *4* = 0.08, *5* = 0.55, *6* = 0.93, *7* = 1.42) and Information Criteria (*AICC* = 10493.1, *BIC* = 10501.9).

This model contained random intercepts for participant-raters (σ^2 = 0.73, *SE* = 0.29) and items (σ^2 = 0.69, *SE* = 0.35). However, when random slopes for items were included, to give a maximal random effects' model, *Retention Interval* was removed each time: immediate and 3-4 hour [*B* = -0.45, *SE(B)* = 0.35, *p* = .194], 2 days to combined previous intervals [*B* = -0.32, *SE(B)* = 0.37, *p* = .388] and 1 week to all previous delays [*B* = -0.51, *SE(B)* = 0.42, *p* = .222]. (This outcome, as observed by Erickson et al., 2022, is mentioned above.)

Experiment 2

(a) Correct Naming

The interaction (*p* = 1.0, *1/Exp(B)* = 3.02) in a full-factorial model emerged greater than alpha and was removed. The subsequent, final model [*F*(2, 397) = 17.67, *p* < .001] comprised both *Early Recall* [*F*(1,397) = 28.83, *p* < .001], with Early Recall > No Early Recall [*B* = 1.52, *SE(B)* = 0.28, *Exp(B)* = 4.57 (2.62, 7.96)]; and *Interview Type* [*F*(1,397) = 9.67, *p* = .002], with H-CI > CI [*B* = 0.84, *SE(B)* = 0.27, *Exp(B)* = 2.33 (1.36, 3.97)]. This model contained random intercepts for items (σ^2 = 3.33, *SE* = 1.74) only; other model details were Overall Correct Classification (82.5%), Intercept [*B* = -2.12, *SE(B)* = 0.64] and Information Criteria (*AICC* = 2063.9, *BIC* = 2067.9).

(b) Mistaken Naming

As for Correct Naming, the interaction in a full-factorial model was removed (*p* = .366, *1/Exp(B)* = 1.93). The subsequent model contained *Interview Type* (*p* = .031) and *Early*

Recall (p = .146, *1/Exp(B)* = 1.68), with the latter emerging greater than alpha, and was removed. The final model comprised *Interview Type* only [$F(1, 398) = 4.44$, p = .036], with H-CI > CI [B = 0.76, *SE(B)* = 0.36, *Exp(B)* = 2.13 (1.05, 4.31)]. The model contained random intercepts for both participant-namers (σ^2 = 0.51, *SE* = 0.28) and items (σ^2 = 0.91, *SE* = 0.53); other details were Overall Correct Classification (84.0%), Intercept [B = -2.04, *SE(B)* = 0.41] and Information Criteria (*AICC* = 1967.4, *BIC* = 1975.3).

(c) Likeness Ratings

The initial, factorial model retained the interaction [$F(1,701) = 41.82$, $p < .001$]; IVs were *Early Recall* [$F(1,701) = 30.04$, $p < .001$] and *Interview Type* [$F(1,701) = 91.72$, $p < .001$]. The interaction, assessed by Fixed Coefficients, revealed that Early Recall was greater than No Early Recall following both CI [B = 0.48, *SE(B)* = 0.20, p = .015, *Exp(B)* = 1.62 (1.10, 2.40)] and H-CI [B = 2.35, *SE(B)* = 0.21, $p < .001$, *Exp(B)* = 10.49 (6.87, 15.98)]; H-CI was greater than CI following early recall [B = 1.72, *SE(B)* = 0.21, $p < .001$, *Exp(B)* = 5.58 (3.73, 8.35)], but there was no difference without early recall [B = -0.15, *SE(B)* = 0.20, p = .47, 1/*Exp(B)* = 1.16 (0.78, 1.72)]. This (final) model included random intercepts for items (σ^2 = 1.90, *SE* = 0.93) only; Threshold rating values of B (*1* = -0.80, *2* = 0.99, *3* = 2.12, *4* = 3.35) and Information Criteria (*AICC* = 9934.4, *BIC* = 9939.0). These tests support the results by GEE.

Next, random slopes were added, to give a maximal random effects' model (incl. random slopes for *Early Recall* for participant-raters and items, *Interview Type* and the interaction for items, and random intercepts for participant-raters and items). The interaction was retained in the model (p = .008), and two of the comparisons still influenced the DV: the benefit of (i) Early Recall following H-CI [p = .004, *SE(B)* = 1.55], and (ii) Interview following Early

Recall [$p < .001$, *SE(B)* = 0.99]; however, there was no longer a benefit of Early Recall following CI [$p = .51$, *SE(B)* = 1.54]. Similar to Experiment 1, adding random slopes increases *SE,* reduces statistical power but does give a more accurate account.

Experiment 3

(a) Correct Naming

The GLMM retained *Early Recall* [$F(1,239) = 3.75$, $p = .054$]: Early Recall marginally > No Early Recall [*B* = 0.73, *SE(B)* = 0.38, *Exp(B)* = 2.07 (0.99, 4.35)]. It contained random intercepts for participant-namers ($\sigma^2 = 0.31$, *SE* = 0.26) and items ($\sigma^2 = 1.72$, *SE* = 0.98); other details were Overall Correct Classification (78.8%), Intercept [*B* = -0.93, *SE(B)* = 0.50] and Information Criteria (*AICC* = 1109.3, *BIC* = 1116.2).

(b) Mistaken Naming

Early Recall was removed [$F(1,239) = 0.46$, $p = .497$]. This means that Early Recall was equivalent to No Early Recall [*B* = -0.21, *SE(B)* = 0.31, *1/Exp(B)* = 1.24 (0.67, 2.28)]. The model contained random intercepts for participant-namers ($\sigma^2 = 0.10$, *SE* = 0.18) and items ($\sigma^2 = 1.06$, *SE* = 0.61); other details were Overall Correct Classification (71.4%), Intercept [*B* = -0.06, *SE(B)* = 0.40] and Information Criteria (*AICC* = 1075.2, *BIC* = 1082.1).

(c) Likeness Ratings

The model comprised *Early Recall* [$F(1,352) = 20.27$, $p < .001$], with Early Recall > No Early Recall [*B* = 0.87, $p < .001$, *SE(B)* = 0.19, *Exp(B)* = 2.37 (1.63, 3.47)]. This model contained random intercepts for both participant-raters ($\sigma^2 = 0.79$, *SE* = 0.33) and items ($\sigma^2 =$ 0.39, *SE* = 0.23); other details were threshold rating values of *B* (*1* = -1.67, *2* = -0.50, *3* = 0.35, *4* = 1.35, *5* = 2.80) and Information Criteria (*AICC* = 5437.8, *BIC* = 5445.5). A

subsequent model with maximal random effects (incl. both random slopes for items and random intercepts for participant-raters) found that the odds of rated likeness was only marginally higher for Early Recall [p = .10, *SE(B)* = 0.62, *Exp(B)* = 2.78 (0.999, 7.71)].

Experiment 4

(a) Correct Naming

The model comprised *Interview Type* [$F(2,256) = 17.84$, $p < .001$]. Fixed Coefficients revealed differences in odds: H-CI > CI [B = 0.93, *SE(B)* = 0.36, p = .010, *Exp(B)* = 2.54 (1.25, 5.16)], ER-H-CI > CI [B = 2.37, *SE(B)* = 0.40, $p < .001$, *Exp(B)* = 10.67 (4.87, 23.39)], and ER-H-CI > H-CI [B = 1.44, *SE(B)* = 0.37, $p < .001$, *Exp(B)* = 4.20 (2.02, 8.75)]. The model contained random intercepts for items (σ^2 = 1.83, *SE* = 1.04) only; other details were Overall Correct Classification (84.9%), Intercept [B = -1.27, *SE(B)* = 0.51] and Information Criteria (*AICC* = 1237.3, *BIC* = 1240.8).

(b) Mistaken Naming

The model contained *Interview Type* [$F(2,256) = 2.57$, p = .078]. The variance of random intercepts was zero for participant-namers (σ^2 = 0.0) and items (σ^2 = 0.0); other details were Overall Correct Classification (89.2%), Intercept [B = -1.68, *SE(B)* = 0.30] and Information Criteria (*AICC* = 1340.2, *BIC* = 1347.2).

Inability to estimate any random effects (here, random intercepts for both participant-namers and items) produced a model where the Hessian matrix is not *positive definite*—that is, it does not converge properly and its validity is uncertain. See Gill and King (2004) for a discussion on this issue. This situation has arisen since mistaken observations were infrequent (N = 3) for composites created in the best condition of the experiment (ER-H-CI),

presumably as these images were constructed very accurately. The consequence was insufficient variability for the model to be able to estimate either random effects. Solutions to this issue include collecting more data (resulting in an increase in total event responses in ER-H-CI), collapsing over conditions (to increase total event responses in the combined category), or to use a generalised-linear but not mixed models (as the random effect of participant responses are taken into account by collapsing over participants or items), as done elsewhere (using GEE) in the paper.

(c) Likeness Ratings

The model included *Interview Type* [$F(2,467) = 87.5$, $p < .001$]: H-CI > CI [$B = 1.52$, $SE(B) = 0.25$, $p < .001$, $Exp(B) = 4.56$ (2.79, 7.47)], ER-H-CI > CI [$B = 3.73$, $SE(B) = 0.29$, $p < .001$, $Exp(B) = 41.56$ (23.71, 72.87)] and ER-H-CI > H-CI [$B = 2.21$, $SE(B) = 0.24$, $p < .001$, $Exp(B) = 9.11$ (5.66, 14.66)]. This model contained random intercepts for both participant-raters ($\sigma^2 = 0.26$, $SE = 0.15$) and items ($\sigma^2 = 0.02$, $SE = 0.05$) (i.e., the variance of random slopes was zero); other details were Threshold rating values of $B\ (3 = 1.37, 4 = 2.87)$ and Information Criteria ($AICC = 3568.4$, $BIC = 3576.6$).

Combined Analyses

Each of the following models followed the procedure described in the main paper. Initial models contained random intercepts for participant-witnesses, participant-namers, items (stimuli) and experiment; the final model contained random effects for which sufficient variance could be estimated from the data.

A. Early Recall

(a) Correct Naming

The model comprised *Early Recall* [$F(1,815) = 11.73$, $p < .001$], with Early Recall > No Early Recall with an overall medium effect size [$B = 1.14$, $SE(B) = 0.33$, $Exp(B) = 3.14$ (1.63, 6.06)]. This model contained random intercepts for participant-witnesses ($\sigma^2 = 1.45$, $SE = 0.42$), items ($\sigma^2 = 1.52$, $SE = 0.71$) and experiments ($\sigma^2 = 0.29$, $SE = 0.45$); other details were Overall Correct Classification (81.9%), Intercept [$B = -1.00$, $SE(B) = 0.48$] and Information Criteria ($AICC = 3896.2$, $BIC = 3910.3$).

(b) Mistaken Naming

Early Recall was retained [$F(1,815) = 3.60$, $p = .058$]: No Early Recall marginally > Early Recall with an overall small effect size [$B = 0.57$, $SE(B) = 0.30$, $Exp(B) = 1.77$ (0.98, 3.20)]. It contained random intercepts for participant-witnesses ($\sigma^2 = 0.71$, $SE = 0.27$), participant-namers ($\sigma^2 = 0.25$, $SE = 0.15$), items ($\sigma^2 = 0.59$, $SE = 0.34$) and experiments ($\sigma^2 = 1.43$, $SE = 1.53$); other details were Overall Correct Classification (84.0%), Intercept [$B = -1.19$, $SE(B) = 0.74$] and Information Criteria ($AICC = 3949.2$, $BIC = 3968.0$).

B. Interview Type

(a) Correct Naming

The model contained *Interview Type* [$F(1,569) = 4.42$, $p = .036$]: H-CI > CI with an overall medium effect size [$B = 0.91$, $SE(B) = 0.43$, $Exp(B) = 2.49$ (1.06, 5.85)]. It contained random intercepts for participant-witnesses ($\sigma^2 = 1.81$, $SE = 0.60$) and items ($\sigma^2 = 1.82$, $SE = 0.92$); other details were Overall Correct Classification (83.9%), Intercept [$B = -1.36$, $SE(B) = 0.44$] and Information Criteria ($AICC = 2779.2$, $BIC = 2787.9$).

(b) Mistaken Naming

Interview Type [$F(1,569) = 1.38$, $p = .241$, *Exp(B)* = 1.45] was not retained in the model. Thus, H-CI was equivalent to CI [$B = 0.37$, *SE(B)* = 0.32, *Exp(B)* = 1.45 (0.78, 2.71)]. The model contained random intercepts for participant-witnesses ($\sigma^2 = 0.37$, $SE = 0.25$), participant-namers ($\sigma^2 = 0.29$, $SE = 0.19$) and items ($\sigma^2 = 0.50$, $SE = 0.30$); other details were Overall Correct Classification (84.6%), Intercept [$B = -1.88$, *SE(B)* = 0.28] and Information Criteria (*AICC* = 2745.6, *BIC* = 2758.6).

(c) Likeness Ratings.

In the previous analyses, power was sufficient for analyses of correct naming and, with the exception of Experiment 2, mistaken naming. In this part, to increase unexpected low statistical power, we also conducted GLMM analyses for likeness ratings across experiments. This proceeded for predictors *Early Recall* (for Experiments 2-3, and then 2-4) and *Interview Type* (Experiments 2 and 4). We followed the same procedure as described above, including use of condensed rating scales, and presenting the model with maximal random effects. The only notable change was that models could now include random intercepts for experiments. In each of the following analyses of combined data, it is apparent that doubling the sample size allowed both predictors to emerge significant. As for analyses of combined naming, GLMMs also included random intercepts for participant-witnesses.

A. Early Recall (Early Recall vs. No Early Recall)

(a) Experiments 2 - 3

The model contained *Early Recall* [$F(1,1060) = 6.31$, $p = .012$]: Early Recall > No Early Recall [$B = 1.14$, *SE(B)* = 0.56, *Exp(B)* = 4.11 (1.36, 12.38)]. The model contained random

intercepts for participant-raters (σ^2 = 0.89, *SE* = 0.26), items (σ^2 = 0.06, *SE* = 0.75) and experiments (σ^2 = 1.47, *SE* = 2.38), and random slopes for *Early Recall* for items (σ^2 = 3.00, *SE* = 1.04); other details were Threshold rating values of *B* (*1* = -1.75, *2* = 0.19, *3* = 1.52, *4* = 2.85, *5* = 5.32) and Information Criteria (*AICC* = 23717.2, *BIC* = 23737.0).

(b) Experiments 2 - 4

The model comprised *Early Recall* [*F*(1,1374) = 14.49, *p* < .001]: Early Recall > No Early Recall [*B* = 1.57, *SE(B)* = 0.41, *Exp(B)* = 4.83 (2.15, 10.87)]. It contained random intercepts for participant-raters (σ^2 = 0.76, *SE* = 0.19), items (σ^2 = 0.07, *SE* = 0.49) and experiments (σ^2 = 2.26, *SE* = 2.44), and random slopes for *Early Recall* for items (σ^2 = 2.37, *SE* = 0.67); other details were Threshold rating values of *B* (*1* = -2.55, *2* = -0.65, *3* = 0.99, *4* = 2.37, *5* = 5.60) and Information Criteria (*AICC* = 34477.5, *BIC* = 34498.4).

B. Interview Type (H-CI vs. CI)

(a) Experiments 2 and 4

The model contained *Interview Type* [*F*(1,1017) = 12.25, *p* < .001]: H-CI > CI [*B* = 0.82, *SE(B)* = 0.24, *Exp(B)* = 2.28 (1.44, 3.62)]. It contained random intercepts for items (σ^2 = 0.95, *SE* = 0.41) and experiments (σ^2 = 1.48, *SE* = 2.27), and random slopes for Interview for items (σ^2 = 0.40, *SE* = 0.17); other details were Threshold rating values of *B* (*1* = -2.03, *2* = -0.45, *3* = 1.78, *4* = 2.69) and Information Criteria (*AICC* = 14602.9, *BIC* = 14617.7).

Appendix D: Discussion on Statistical Power, Approach and GLMM

Regarding statistical approach and power, experiments were designed (see Supplementary Materials Sections 4.1.1 – 4.1.2) to be able to detect a practically-useful medium effect [*Exp(B)* ≈ 2.5] by analysis using GEE. Given concern over reduced power when including random effects for participant-witnesses, analyses for correct naming took this random effect into account in a combined analysis across experiments. The approach was effective.

We have since re-run the analyses for each experiment including random intercepts for participant-witnesses. The exercise revealed, as early recall emerged as a medium effect, the same pattern of significant (Experiment 2) or marginally-significant (Experiment 4, see below) results. There were inconsistent results (i.e., between by-participants and by-items analyses) as the effect size was small for Early Recall in Experiment 3, and Interview Type in both Experiment 2 and the combined analysis. Therefore, sample size had been estimated appropriately. We note, though, that, when including random intercepts for participant-witnesses, the marginally-significant result in Experiment 4 (H-CI > CI, $p = .08$) was a consequence of the alpha used for the post-hoc tests; these require $\alpha = .05$ (cf. $\alpha = .1$ to retain predictors in a Model), and so a larger sample would have been appropriate for this experiment—an estimated increase of 58 responses, or three more participant-namers per group.

Participant-namer responses were also analysed using Generalized Linear Mixed Models (Appendices C and E). GLMM is gaining popularity in Psychology (Meteyard & Davies, 2020), and has been used to analyse data from a single-experiment composite paper by Erickson et al. (2022). As a unified model, GLMM has the advantage that a single conclusion can be readily made, unlike GEE. In fact, in Experiment 2, inferential results for GEE turned

out to be inconsistent: following early (cf. no early) recall, the odds of a mistaken name were marginally lower by-items (p = .066), but not significant by-participants (p = .15): by GLMM, this predictor was removed from the model (p = .17), indicating a non-significant effect. Also, in the combined analyses, while GEE led to consistent results for Early Recall, this was not the case for Interview (as the size of the effect was smaller than that planned). Overall, the outcome of GLMM supported the significant and non-significant findings from GEE for the primary DV, correct naming. Results were also consistent by mistaken naming, except that there were insufficient data for mistaken naming in Experiment 4, a situation that presumably has occurred as the composites were very accurate in the best condition, generating infrequent mistaken names. This situation is readily overcome for either type of analysis, such as by collecting more data or collapsing over conditions (see Gill & King, 2004). We note that including a random effect of participant-witness in the individual experiments based on a medium effect size led to the same conclusions as GEE (as discussed in the previous paragraph). For likeness ratings, considerable increase in SE occurred when random slopes were included in the random effects' model, and analyses were shown to benefit from doubling the sample size. So, taking into account the requirement of a greater sample size for analysing likeness ratings, the single inferential outcome provided by GLMM (cf. GEE) suggests greater utility.

Appendix E: Comparison of Analyses for Naming and Likeness for GEE (by-participants and by-items) and GLMM

The following table compares the main inferential statistics conducted for the three methods of analyses by experiment (Expt) and DV (Task).

[Table 15]

Appendix F. Follow-up Experiment involving Early Verbal Recall

We followed the same basic design and procedure as that described in Experiment 4, with *Interview Type* comprising No Early Recall, Early Written Recall (EWR) and Early Verbal Recall (EVR). Participant-witnesses were asked to freely recall the face 3-4 hr after encoding either (i) for EWR, in written format (as done in the experiments so far) or (ii) for EVR, verbally, to the researcher (as in Brown et al., 2017). Materials were 10 characters from Coronation Street, as used in Experiment 2. All 30 participant-witnesses (12 female, 18 male; Age: 18-56, M = 29.4, SD = 13.0 years) received an H-CI prior to EvoFIT face construction, administered 20-28 hr after encoding a single target face. Composite naming was conducted by 63 participant-namers (29 female, 34 male; Age: 18-56, M = 30.8, SD = 12.9 years). Participant-witnesses and -namers were opportunity sampled from staff, students, and members of the public (and coding for these random variables, along with for items, were included in the analyses).

For correct naming, GEE, by-participants, retained *Interview Type* (1 = *H-CI*, 2 = *EWR+H-CI*, 3 = *EVR+H-CI*) [$\chi_1^2(2) = 11.31$, $p = .004$]. Relative to No Early Recall, while Early Written Recall led to odds of a correct response that was higher [$p = .002$, $Exp(B) = 2.68$], composite naming did not benefit from Early Verbal Recall [$p = .63$, $Exp(B) = 1.18$]. Model parameters: Intercept [$B = -1.11$, $SE(B) = 0.25$] and Information Criteria ($QIC = 790.7$, $QICC = 782.4$). For mistaken naming, with respect to No Early Recall, while means were somewhat lower for both EWR ($MD = 10.1\%$) and EVR ($MD = 13.5\%$), *Interview Type* was not retained in the model [$\chi_1^2(2) = 2.38$, $p = .304$, $1/Exp(B) = 1.52 – 1.76$].

GEE By-items: The conclusions reached were the same. For correct naming, the model retained *Interview Type* [$\chi_1^2(2) = 29.13$, $p < .001$]: Early Written Recall > No Recall [$p < .001$, *Exp(B)* = 2.68] and Early Verbal Recall = No Early Recall [$p = .43$, *Exp(B)* = 1.18]. Intercept [$B = -1.11$, *SE(B)* = 0.28] and Information Criteria (*QIC* = 798.8, *QICC* = 782.4). For mistaken naming, *Interview Type* was retained [$\chi_1^2(2) = 8.84$, $p = .012$]: No Early Recall > Early Written Recall [*Exp(B)* = 1.52, $p = .037$] and No Early Recall > Early Verbal Recall [*Exp(B)* = 1.76, $p = .004$].

GLMM: Conclusions were also the same. For correct naming, *Interview Type* was retained [$\chi_1^2(2,627) = 5.45$, $p = .004$]: Early Written Recall > No Early Recall [$p = .003$, *Exp(B)* = 3.57] and Early Verbal Recall = No Early Recall [$p = .69$, *Exp(B)* = 1.19]. The model contained random intercepts for participant-witnesses ($\sigma^2 = 1.27$, *SE* = 0.37) and items ($\sigma^2 = 1.21$, *SE* = 0.67), Other model details were Overall Correct Classification (83.5%), Intercept [$B = -1.48$, *SE(B)* = 0.47] and Information Criteria (*AICC* = 3017.6, *BIC* = 3026.5). For mistaken naming, *Interview Type* was not retained in the model [$\chi_1^2(2,627) = 1.12$, $p = .328$, *1/Exp(B)* = 1.68 – 1.94].

Appendix G. Early Verbal Recall following Longer Retention

We tested the suggestion that early written recall would still be effective after a longer, nominal 24-hr (cf. 3-4 hr previously) retention interval, with all participant-witnesses constructing composites 48-hr after encoding. Two factors were manipulated, *Early Recall* (*0* = No Early Recall, *1* = Early Written Recall) and *Interview Type* (*0* = CI, *1* = H-CI), in a 2 × 2 between-subjects full-factorial design. Both factors were implemented as described in the paper. The target identities were 10 male footballers playing at international level in the UK. Face construction was carried out by 40 participant-witnesses (12 female, 28 male; Age: 18-75, M = 30.0, SD = 14.4 years). Following randomisation, half of these participants were given the instruction, as before, to write down a detailed description of the face (independently, 20-28 hours after encoding). All participants constructed the face using EvoFIT between 44 and 52 hours after encoding, following a CI or an H-CI. Composite face construction was carried out remotely, using a self-directed procedure where participants followed instructions presented on the computer screen. Composite naming was also carried out remotely, by 44 participants, with equal sampling (9 female, 35 male; Age: 18-62, M = 31.0, SD = 11.8 years). Participant-witnesses and -namers were an opportunity sample comprising students at the University of Lancashire and members of the public. As before, participant-witnesses and -namers were included as random effects, along with items, in all analyses. All three analyses produced the same pattern of significant, marginal and non-significant differences.

For correct naming, in a full factorial model, GEE, by-participants, the interaction [p = .668, *1/Exp(B)* = 1.22], was greater than alpha and was removed. The resulting, final model comprised *Early Recall* [$\chi^2(1) = 12.64$, $p < .001$], as Early Written Recall > No Early Recall

[*Exp(B)* = 2.23], and H-CI > CI [*Exp(B)* = 1.86]. Other model details were Intercept [*B* = -1.75, *SE(B)* = 0.22] and Information Criteria (*QIC* = 492.4, *QICC* = 492.4). For mistaken naming, the interaction [*p* = .337, *Exp(B)* = 1.51] was removed from the model, as were both IVs when tested together in the subsequent model [*p*s > .19, *1/Exp(B)* = 1.11-1.32].

GEE By-items: For correct naming, the interaction in a full-factorial model was removed [*p* = .777, *1/Exp(B)* = 1.21]. The resulting model contained *Early Recall* [$\chi^2(1) = 6.25$, *p* = .015] with a benefit for early recall [*Exp(B)* = 2.23]; and *Interview Type* [$\chi^2(1) = 3.74$, *p* = .057] with a marginal benefit for H-CI [*Exp(B)* = 1.86]; Intercept [*B* = -1.76, *SE(B)* = 0.32] and Information Criteria (*QIC* = 499.6, *QICC* = 496.4). For mistaken naming, the interaction [*p* = .432, *Exp(B)* = 1.51] was removed in the full-factorial model, as were both individual predictors tested together in the subsequent model [*p*s > .29, *1/Exp(B)* = 1.11-1.32].

GLMM: For correct naming, in a full-factorial model, the interaction [*p* = .774, *1/Exp(B)* = 1.21] was greater than alpha and was removed. The resulting, final model retained both *Early Recall* [$F(1, 430) = 6.73$, *p* = .010] and *Interview Type* [$F(1, 430) = 4.37$, *p* = .037]: Early Written Recall > No Early Recall [*Exp(B)* = 2.29], and H-CI > CI [*Exp(B)* = 1.95]. The final model included random intercepts for participant-witnesses ($\sigma^2 = 0.43$, *SE* = 0.27) and items ($\sigma^2 = 0.46$, *SE* = 0.38). Other details were Overall model [$F(2,430) = 5.30$, *p* = .005], Overall Correct Classification (78.3%), Intercept [*B* = -1.88, *SE(B)* = 0.37] and Information Criteria (*AICC* = 2011.9, *BIC* = 2020.0). For mistaken naming, the interaction [*p* = .432, *Exp(B)* = 1.49] was removed; both individual predictors were also removed when tested together in the subsequent model [*p*s > .23, *1/Exp(B)* = 1.12-1.34].

Figures

Figure 1: Example composites constructed to resemble the footballer Steven Gerrard. Each picture was created by a different person after experiencing one of four retention intervals from encoding, from left-to-right: immediate, 3-4 hours, 2 days and 1 week. For reasons of copyright, the actual target picture cannot be reproduced; however, a photograph (far right) of this player, taken around the same time, was located on Wikimedia Commons (note that the image used in the project was a more frontal view of the face).

Alt Text for figure: The figure shows a colour photograph of the UK footballer Steven Gerrard on the far right. The four greyscale images on the left show facial composites that were constructed to resemble this individual at each retention interval used in the experiment (immediate, 3-4 hours, 2 days and 1 week).

Table 1. Procedural Flowchart for Stages 1 – 3 of all experiments (with reference to relevant sections of the Online Supplementary Materials)

Experimental Stage	Participant Tasks	Online Supplementary Materials reference
Stage 1: composite construction (participant-witnesses)	Part 1: Unfamiliar target viewing	Section 1.1: Materials and Procedure
	Part 2: Self-administered written early-recall interview[a]	Section 1.2: Materials and Procedure
	Part 3:	
	1. Pre-construction Cognitive Interview (all experiments)	Section 1.3: Materials and Procedure
	2. Holistic-Cognitive Interview (Experiments 2 and 4)	Section 1.4.1: Materials and Procedure
	3. Modified eye-region Holistic-Cognitive Interview (Experiment 4, only)	Section 1.4.2: Materials and Procedure
	4. Composite construction[b], using:	
	i) PRO-fit (Experiments 1 - 2)	Section 1.5.1: Procedure
	ii) sketch (Experiment 3)	Section 1.5.2: Procedure
	iii) EvoFIT (Experiment 4).	Section 1.5.3: Procedure
Stage 2: Composite Naming (participant-namers)	Composite and target photograph naming (all experiments)[c].	Section 2.1: Procedure and Materials Section 4.1.1: Power and Inferential Analyses
Stage 3: Composite Evaluation (participant-raters)	Composite Likeness Rating (all experiments)[d]	Section 3.1: Procedure and Materials Section 4.1.2: Power and Inferential Analysis

Note. a) The self-administered written early-recall interview was used in Experiments 2 – 4, only, and occurred 3 – 4 hours after target encoding. b) Composite construction always occurred 1 day after unfamiliar-target encoding in Experiments 2 – 4, but at variable retention intervals in Experiment 1 (immediately, after 3 – 4 hours, after 1 – 2 days, or after 1 week). c) Stage 2 (composite and target photograph naming) was always completed by target-familiar participants. d) Stage 3 (composite likeness ratings) was always completed by target-unfamiliar participants.

Table 2. Experiment 1 Results for Each DV (Correct Naming, Mistaken Naming and Likeness Ratings) by Increasing *Retention Interval*.

DV	*Retention Interval* [1]			
	Immediate	3 – 4 hours	2 days	1 week
Correct Naming	27.1 (26 / 96)	12.4 (12 / 97)	9.0 (9 / 100)	3.1 (3 / 97)
Mistaken Naming	29.2 (28 / 96)	26.8 (26 / 97)	34.0 (34 / 100)	53.6 (52 / 97)
Likeness Ratings	4.5 (0.2)	4.0 (0.2)	4.0 (0.2)	3.5 (0.2)

Note. *Correct Naming*: Shown as percentage, and (underneath) as number of correct names offered (numerator) out of the number of correctly identified targets (denominator). [1] $p < .001$. *Mistaken Naming*: Shown as percentage, and (underneath) as number of mistaken names offered (numerator) out of the number of correctly identified targets (denominator). [1] $p < .02$. *Likeness Ratings*: Rating scale (1 = *very dissimilar* ... 15 = *very alike*; with scale points 9 – 15 recoded as 8). Shown are mean values and (underneath) *SE*. [1] $p < .01$. For all analyses, results are specified with respect to the lowest category, underlined (here, Immediate); predictors were sorted in descending order; target (DV) were sorted in descending order (except likeness, ascending); see Appendix A for associated statistics, Appendix B for analyses by-items, Appendix C for analyses by GLMM and Appendix E for table of statistical comparisons.

Table 3. Percentage Correct Naming of PRO-fit Composites Constructed by *Early Recall* and *Interview Type*.

	Early Recall [1]		
Interview Type [2]	No Early Recall	Early Recall	Mean
CI	20.0 (20 / 100)	40.0 (40 / 100)	30.0 (60 / 200)
H-CI	30.0 (30 / 100)	54.0 (54 / 100)	42.0 (84 / 200)
Mean	25.0 (50 / 200)	47.0 (94 / 200)	36.0 (144 / 400)

Note. [1,2] $p < .001$.

Table 4. Model Parameters for the Impact of *Early Recall* and *Interview Type* on Correct Naming of PRO-fit Composites.

	B	*SE(B)*	$\chi_1^2(1)$	*p*	*Exp(B)*	*95% CI(-)*	*95% CI(+)*
Fixed Effects							
Early Recall vs. No Early Recall	1.00	0.13	61.51	< .001	2.71	2.11	3.47
H-CI vs. CI	0.55	0.13	19.36	< .001	1.74	1.36	2.23

Note. For the by-participants analysis, Fixed Effects (IVs) presented are coefficients [*B*], standard error [*SE(B)*], model fit [χ_1^2 and *p*] and corresponding Odds Ratio [*Exp(B)*]; Model Intercept [*B* = -1.40, *SE(B)* = 0.12]. Based on Cohen's (1988) estimates, an odds ratio of around 1.5 can be considered a "small" effect size, 2.5 as "medium" and 4.5 as "large" (Sporer & Martschuk, 2014). For example, an odds ratio of 2.71 is therefore a medium effect, and means that the odds of a correct name following early recall is 2.71 times the odds of a correct name with no early recall. See Appendix B for by-items analysis.

Table 5. Percentage Mistaken Naming of PRO-fit Composites by *Interview Type*.

	Early Recall		
Interview Type [1]	No Early Recall	Early Recall	Mean
CI	20.0 (20 / 100)	10.0 (10 / 100)	15.0 (30 / 200)
H-CI	27.0 (27 / 100)	23.0 (23 / 100)	25.0 (50 / 200)
Mean	23.5 (47 / 200)	16.5 (33 / 200)	20.0 (80 / 400)

Note. Early Recall was removed from the model by-participants (p = .15, *1/Exp(B)* = 1.56) but was retained with the IV marginally-significant by-items (p = .066, *1/Exp(B)* = 1.56, Appendix B), consistent with the emerging small effect. The final Model comprised *Interview Type*: H-CI > CI [B = 0.64, *SE(B)* = 0.32, *Exp(B)* = 1.89 (1.02, 3.51)]; Intercept [B = -1.74, *SE(B)* = 0.24]. [1] $p < .05$.

Table 6. Mean Likeness Ratings (*SE*) of PRO-fit Composites Constructed by *Early Recall* and *Interview Type*.

	Early Recall	
Interview Type	No Early Recall	Early Recall
CI	2.2[a] (0.1)	2.5 (0.1)
H-CI	2.1[a] (0.1)	3.4 (0.1)

Note. Rating scale (1 = *very poor likeness* … 7 = *very good likeness*; with scale points 6 and 7 recoded as 5). The interaction indicated inconsistent odds between *Early Recall* and *Interview Type* ($p < .001$): Early Recall > No Early Recall: CI [$B = 0.54$, $SE(B) = 0.19$, $p = .005$, $Exp(B) = 1.72$ (1.18, 2.52)] and H-CI [$B = 1.70$, $SE(B) = 0.20$, $p < .001$, $Exp(B) = 5.49$ (3.73, 8.06)]. H-CI > CI: Early Recall [$B = 1.16$, $SE(B) = 0.19$, $p < .001$, $Exp(B) = 3.18$ (2.18, 4.63)] and No Early Recall (ns) [$B = 0.001$, $SE(B) = 0.20$, $p = 1.0$, $Exp(B) = 1.001$ (0.63, 1.47)]. All pairwise comparisons were significant ($ps \leq .005$) except [a] $p = 1.0$.

Table 7. Percentage Correct Naming of Sketch Composites Constructed by *Early Recall*.

Early Recall [1]	
No Recall	Recall
34.5 (39 / 113)	47.7 (61 / 128)

Note. [1] $p < .05$.

Table 8. Model Parameters for the Impact of *Early Recall* on Correct Naming for Sketch Composites.

	B	*SE(B)*	$\chi_1^2(1)$	*p*	*Exp(B)*	*95% CI(-)*	*95% CI(+)*
Fixed Effects							
Early Recall vs. No Early Recall	0.55	0.27	4.08	.043	1.73	1.02	2.94

Note. Model Intercept [*B* = -0.64, *SE(B)* = 0.20].

Table 9. Mean Likeness Ratings (*SE*) of Sketch Composites by *Early Recall*.

Early Recall [1]	
No Early Recall	Early Recall
3.2 (0.1)	3.8 (0.1)

Note. Rating scale (1 = *very poor likeness* … 7 = *very good likeness;* with scale point 7 recoded as 6). Early Recall > No Early Recall [*B* = 0.69, *SE(B)* = 0.17, *Exp(B)* = 1.99 (1.42, 2.79)]. [1] $p < .001$.

Table 10. Percentage Correct Naming of EvoFIT Composites Constructed by *Interview Type* (CI vs. H-CI vs. ER-H-CI).

Interview Type [1]		
CI	H-CI	ER-H-CI
28.9 (24 / 83)	45.5 (40 / 88)	71.6 (63 / 88)

Note. [1] $p < .001$: all comparisons, $p < .001$.

Table 11. Model Parameters for the Impact of *Interview Type* (CI vs. H-CI vs. ER-H-CI) on Correct Naming of EvoFIT Composites.

	B	*SE(B)*	$\chi_1^2(1)$	*p*	*Exp(B)*	*95% CI(-)*	*95% CI(+)*
Fixed Effects							
H-CI vs. CI	0.70	0.19	13.64	< .001	2.02	1.39	2.94
ER-H-CI vs.CI	1.80	0.20	81.52	< .001	6.03	4.08	8.91
ER-H-CI vs. H-CI	1.09	0.18	35.30	< .001	2.98	2.08	4.28

Note. Model Intercept [B = -0.89, $SE(B)$ = 0.15].

Table 12. Percentage Mistaken Naming of EvoFIT Composites Constructed by *Interview Type* (CI vs. H-CI vs. ER-H-CI).

Interview Type [1]		
CI	H-CI	ER-H-CI
15.7 (13 / 83)	13.6 (12 / 88)	3.4 (3 / 88)

Note. CI = H-CI (ns) [$B = 0.17$, $SE(B) = 0.41$, $p = .85$, $Exp(B) = 1.18$ (0.53, 2.61)], CI > ER-H-CI [$B = 1.66$, $SE(B) = 0.62$, $p = .007$, $Exp(B) = 5.26$ (1.57, 17.61)] and H-CI > ER-H-CI [$B = 1.49$, $SE(B) = 0.62$, $p = .016$, $Exp(B) = 4.45$ (1.32, 15.02)]. Intercept [$B = -1.68$, $SE(B) = 0.28$]. [1] $p < .05$.

Table 13. Mean Likeness Ratings (*SE*) of EvoFIT Composites Constructed by *Interview Type* (CI vs. H-CI vs. ER-H-CI).

Interview Type [1]		
CI	H-CI	ER-H-CI
3.3 (0.04)	3.8 (0.06)	4.6 (0.05)

Note. Rating scale (1 = *very poor likeness* … 7 = *very good likeness*; with scale points 1 and 2 recoded as 3, and *6* and *7* recoded as *5*). H-CI > CI [$B = 1.48$, $SE(B) = 0.24$, $p < .001$, $Exp(B) = 4.39$ (2.72, 7.08)], ER-H-CI > CI [$B = 3.60$, $SE(B) = 0.28$, $p < .001$, $Exp(B) = 36.76$ (21.10, 64.04)] and ER-H-CI > H-CI [$B = 2.13$, $SE(B) = 0.24$, $p < .001$, $Exp(B) = 8.40$ (5.26, 13.33)]. [1] $p < .001$; all comparisons, $p < .001$.

Table 14. Means for each DV (Correct Naming, Mistaken Naming and Likeness Rating) by Composite System and Experiment.

DV	*Interview Technique*			
	CI	ER-CI	H-CI	ER-H-CI
Correct Naming				
PRO-fit (Experiment 2)	20.0	40.0	30.0	54.0
Sketch (Experiment 3)	34.5	47.7		
EvoFIT (Experiment 4)	28.9		45.5	71.6
Mistaken Naming				
PRO-fit (Experiment 2)	20.0	10.0	27.0	23.0
Sketch (Experiment 3)	43.8	47.8		
EvoFIT (Experiment 4)	15.7		13.6	3.4
Likeness Rating				
PRO-fit (Experiment 2)	2.2	2.5	2.1	3.4
Sketch (Experiment 3)	3.2	3.8		
EvoFIT (Experiment 4)	3.3		3.8	4.6

Note. CI = face-recall CI, ER-CI = early recall + face recall CI, H-CI = face and holistic recall, and ER-H-CI = early recall + face and holistic recall. Values are expressed in percentages for Correct Naming and Mistaken Naming, and using the mean for Likeness Rating.

1

Table 15. Comparison of Analyses for Naming rates and Likeness Ratings for GEE (by-participants and by-items) and GLMM, by Experiment (Expt) and Dependent Variable (Task)

			GEE (by-participants)				GEE (by-items)				GLMM			
Expt	Task	Predictor	χ_1^2	*p*	*SE*	*Exp(B)*	χ_2^2	*p*	*SE*	*Exp(B)*	*F*	*p*	*SE*	*Exp(B)*
1	Correct Naming	Retention Interval	39.13	< .001	-	-	20.35	< .001	-	-	7.52	< .001	-	-
1	Correct Naming	First Contrast	-	< .001	0.04	2.67	-	.008	0.06	2.63	3.20	.016	0.40	2.60
2	Correct Naming	Early Recall	61.51	< .001	0.13	2.71	28.71	< .001	0.19	2.71	28.83	< .001	0.28	4.57
2	Correct Naming	Interview Type	19.36	< .001	0.13	1.74	9.69	< .001	0.18	1.74	9.67	.002	0.27	2.33
3	Correct Naming	Early Recall	4.08	.043	0.27	1.73	5.77	.016	0.24	1.76	3.75	.054	0.38	2.07
4	Correct Naming	Interview Type	80.03	< .001	-	-	38.90	< .001	-	-	17.84	< .001	-	-
4	Correct Naming	ER-H-CI > H-CI	-	< .001	0.18	2.98	-	< .001	0.29	3.01	-	< .001	0.37	4.20
4	Correct Naming	H-CI > CI	-	< .001	0.19	2.02	-	.008	0.29	2.17	-	.010	0.36	2.54
2-4	Correct Naming	Early Recall	33.67	< .001	0.15	2.32	7.49	.006	0.31	2.36	11.73	< .001	0.33	3.14
2+4	Correct Naming	Interview Type	10.93	< .001	0.18	1.79	2.30	.130	0.39	1.82	4.42	.036	0.43	2.49
1	Mistaken naming	Retention Interval	10.80	.013	-	-	18.44	< .001	-	-	6.11	< .001	-	-
1	Mistaken naming	First Contrast	-	.750	0.08	-	-	.710	0.06	-	-	.710	0.32	-
2	Mistaken naming	Early Recall	2.07	.150	0.31	1.56†	3.40	.066	0.24	1.56†	1.88	.170	0.36	1.64†
2	Mistaken naming	Interview Type	4.05	.044	0.32	1.89	6.66	.010	0.25	1.89	4.44	.036	0.36	2.13
3	Mistaken naming	Early Recall	0.41	.410	-	-	0.55	.460	-	-	0.46	.500	-	-
4	Mistaken naming	Interview Type	7.47	.024	-	-	6.52	.038	-	-	2.57*	.078	-	-
4	Mistaken naming	H-CI > ER-H-CI	-	.016	0.62	4.45	-	.024	0.66	4.48	-	.013	0.53	2.26
4	Mistaken naming	CI > ER H-CI	-	.007	0.62	5.26	-	.012	0.66	5.26	-	.027	0.58	3.62
2-4	Mistaken naming	Early Recall	3.98	.046	0.15	1.38†	1.22	.270	0.34	1.45†	3.60	.058	0.30	1.77†
2+4	Mistaken naming	Interview Type	3.78	.052	0.22	1.53	1.63	.200	0.32	1.51	1.38	.240	0.32	1.45
1	Likeness Rating	Retention Interval	12.36	.006	-	-	10.13	.018	-	-	1.26	.290	-	-
1	Likeness Rating	First Contrast	-	.085	0.18	1.37	-	.079	0.18	1.38	-	.190	0.35	1.57
2	Likeness Rating	Early Recall	65.31	< .001	-	-	67.55	< .001	-	-	3.76	.053	-	-
2	Likeness Rating	Early Recall: CI	-	.005	0.19	1.72	-	.007	0.16	1.53	-	.510	1.54	2.78
2	Likeness Rating	Early Recall H-CI	-	< .001	0.20	5.49	-	< .001	0.19	5.26	-	.004	1.55	8.33
2	Likeness Rating	Interview Type	17.94	< .001	-	-	16.65	< .001	-	-	4.50	.053	-	-
2	Likeness Rating	Interview: Early Recall	-	< .001	0.19	3.18	-	< .001	0.18	2.97	-	< .001	0.99	27.03
3	Likeness Rating	Early Recall	15.93	< .001	0.17	1.99	14.32	< .001	0.19	2.02	3.65	.056	0.92	5.81
4	Likeness Rating	Interview Type	166.13	< .001	-	-	38.90	< .001	-	-	87.46	< .001	-	-
4	Likeness Rating	H-CI > CI	-	< .001	0.24	4.39	-	< .001	0.22	4.28	-	< .001	0.25	4.56
4	Likeness Rating	ER-H-CI > H-CI	-	< .001	0.24	8.40	-	< .001	0.24	8.47	-	< .001	0.24	9.09

† For ease of interpretation, as it is better for this measure of effect size to be greater than 1.0 (Osborne, 2016), the value is expressed as the exponential of the absolute value of *B* [similar to *1/Exp(B)*, as used in the paper]. In these cases, Early Recall leads to lower mistaken composite naming than No Early Recall.

* Model is not considered valid (since no random effects were able to be estimated); GEE is advised as an alternative technique for analysing this data set (see Appendix C).

Online Supplementary Materials

1. General Method

1.1 Stage 1: Target Encoding (Procedure and Materials, all experiments)

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

To facilitate generalisation of results, different target identity pools were purposely used in each experiment (*see* interim method sections). However, all target photographs were prepared and presented to the same standard across experiments. Specifically, good-quality photographs of each target identity, sourced from the internet, depicted the head and shoulders of the individual, who was adopting a front-facing, neutral pose, with minimal facial hair and no adornments (e.g., no target faces had a nose stud) that might otherwise render the face too distinctive. Per experiment, a copy of these target photographs was prepared in an electronic document for each condition, in colour, at 8 cm width x 10 cm height, one per A4 page. For face-to-face interactions (Experiments 1, 2 and 4), these documents were reproduced using a good quality printer.

[9] 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.

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

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

During the target-viewing session, participants assigned to the early recall condition received a sealed envelope from the experimenter (Experiments 2 and 4). They were told to open the envelope 3-4 hours later, and follow the printed instructions therein, which asked them to write down as much as they could remember about the face on the enclosed A4 sheet of paper (i.e., a free-recall attempt). While participants were not subsequently reminded to complete the task, they were requested to return this description to the experimenter when they attended their next experimental session (described below), as a compliance check[10]. Participants were not required to review this description ahead of their next experimental session [comprising the CI, or (the original or modified) H-CI, and composite construction] as research suggests that reviewing a retrieval attempt does not facilitate subsequent recall (e.g., Sauerland et al., 2008; Turtle & Yuille, 1994).

The procedure for requesting early recall was adapted to be remote for Experiment 3, due to restrictions imposed by the COVID-19 pandemic. Here, 3-4 hours after encoding, the researcher contacted participants assigned to the early-recall condition by telephone, requesting them to write down a description of the target face once the call had ended. In the following meeting, all participants reported that they had completed the exercise, as requested.

1.3 *Stage 1: Practitioner-led Cognitive Interview (Materials and experiment-specific procedures)*

[10] 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.

Participant-witnesses began their final experimental session with completion of a three-stage, face-recall Cognitive Interview (CI), which was conducted online for Experiment 3 (via FaceTime or Skype), and in-person for the other experiments. We describe the interview procedure in Experiment 1 > Method > Procedure. As part of this interview, the experimenter used an A4 paper sheet to write down the participant-witnesses' free and cued recall attempt. The sheet contained sub-headings that referred to each facial region and / or feature: overall facial characteristics, facial shape, hair, eyebrows, eyes, nose, mouth and ears. The cued-recall stage of the interview (where the participant-witnesses' freely-recalled descriptors are repeated back to them, and further recall is prompted by the experimenter) was omitted from Experiment 4, as this mnemonic does not appear to facilitate EvoFIT construction (e.g., Frowd et al., 2015).

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

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

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

In Experiment 4, a third of participant-witnesses were assigned to receive a revised version of the H-CI. For EvoFIT, Skelton et al. (2020) found enhanced composite effectiveness when participants provided the aforementioned holistic ratings twice: once for the whole-face and then again when focusing on the eye region (the area including the eyes and eyebrows). Potentially harnessing Transfer Appropriate Processing (TAP; Morris et al., 1977) mechanisms, this restricted focus aligns with that instructed during EvoFIT array presentation, where witnesses are encouraged to focus on the likeness of the eye-region when making their face selections (Fodarella et al., 2017). Here then, participant-witnesses used the same three-point scale to rate the extent to which they perceived the eye region to convey the same seven characteristics (as above) of the target's character, with the experimenter again recording these ratings on the aforementioned response sheet.

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

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

The experimenter first independently entered the descriptors provided by the participant-witness during the CI, as recorded on the description sheet, to locate approximately 20 'matching' system-housed photographic exemplars, per facial feature (e.g., for the eyes, nose, mouth, etc.). The experimenter then showed the participant the returned exemplar sub-set, per feature, embedded within the context of a whole-face, and the participant was asked to direct the experimenter toward the single best exemplar, per feature category. With these best

feature exemplars in place, the participant was then invited to suggest how the likeness of the face could be improved, with the experimenter using editing tools to re-position, re-size and re-shade facial features, as requested. PRO-fit construction took approximately 1-hour, including debriefing.

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

An established procedure of sketch production (e.g., Fodarella et al., 2015; Frowd et al., 2005) was implemented by an extensively-trained, target-naïve, artist. Due to restrictions imposed by the COVID-19 pandemic, interaction with participant-witnesses was carried out via video link (FaceTime or Skype), a procedure previously found to be effective for construction of forensic sketches (Kuivaniemi-Smith et al., 2014). Directly consulting the participant's face description, obtained during the CI, the artist prepared an initial sketch, wherein facial features were faintly drawn. The artist then followed instructions, given by the participant, to improve image likeness, altering feature size, position and shading. Sketched composites took around two hours to construct, including debriefing.

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

An extensively-trained, target-naïve experimenter controlled the software. The EvoFIT construction process is described in detail elsewhere (e.g., Fodarella et al., 2015), and thus a brief protocol is presented here. Participant-witnesses first directed the experimenter to a database that matched the previously-seen target for age and gender. Participants were then presented with four screens of 18 'smooth' (texture-averaged) faces that revealed the internal-features region (i.e., the facial area excluding hair, forehead, ears and neck): they were asked to ignore face width but indicate to the experimenter the best two matching items from each

of the first three screens, based on the target-likeness of the eye region. The participant-witness could review their selections, and make any replacements, on a fourth screen. This procedure was repeated over four screens of 'textured' faces (presented with variable facial texture), with participants then presented with a combination of previously-chosen smooth and textured faces from which they directed the experimenter towards the single best match. Participants undertook a second experimenter-led iteration, with previous choices combined, to 'evolve' a face. The participant then directed the experimenter to enhance the likeness, first using holistic tools: scales that changed width, weight, age, and 12 further overall properties of the face. The face was then subject to further enhancement: the experimenter could first adjust greyscale shading of features and then feature shape and position on the face. Hair and other external features were added, and the aforementioned software tools were used again, as required, with the aim of creating the best likeness possible. The procedure took approximately 45 minutes, including debriefing.

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

Mirroring the forensic situation, target-familiar participants were recruited to attempt to name the composites produced during Stage 1, with the following procedure conducted in-person for Experiments 1 and 4, and remotely (via FaceTime or Skype) for Experiments 2 and 3. Participant-namers were tested individually, and the task was self-paced. Each participant was randomly allocated to view the composites constructed in only one of the Stage 1 conditions of that experiment, with items presented by the experimenter sequentially, in a different random order for each person. Composites were sized to 8 cm (width) x 10 cm (height) in electronic documents. Each document contained 10 composites (one per target identity), each

presented individually per A4 page, in greyscale, which were printed to good quality for face-to-face interactions. Participants were asked to name each composite, saying a name if one came to mind; otherwise, a "don't know" response was acceptable.

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

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

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

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

Participant-raters tend to judge visual match more harshly for identities with whom they are familiar than unfamiliar (Frowd, 2021), and so target-unfamiliar participants were recruited to

Stage 3. As such, data retention principles contrasted with those implemented in Stage 2: if the participant *did* recognise one or two of the target identities (as assessed via a final photograph naming task, described below), their data for those individual composites were discarded; if they recognised more than two, the participant was replaced, with the latter instance occurring rarely across experiments.

Participant-raters were tested individually, either face-to-face (Experiments 1 and 4), or remotely (Experiments 2 and 3, via FaceTime or Skype) and the task was self-paced. A within-subjects design was adopted: For each target identity, participants were concurrently presented with all of Stage 1's corresponding composites (i.e., one facial image resulting from each construction condition) and the corresponding target photograph. Composite array-to-target photograph slides were presented randomised by target identity and participant, with both composites and target photograph images sized to the same dimensions as in Stage 2. Per composite-to-target pairing, participants were asked to assess the likeness between the two images, with absolute judgments given in Experiment 1 (i.e., participants made a composite-to-target rating for the first composite and first target identity, before viewing and rating the second composite according to its likeness again to the first target identity, and so on, until they had provided a likeness rating for all composites constructed to resemble that target identity). Subsequent experiments instead required relative likeness judgments to be made (i.e., participants first passively viewed all composites constructed to resemble a particular identity before they sequentially rated the likeness between each of those composites and the same target identity). The latter task variation was made as it can be difficult to judge variation in likeness without first inspecting the relevant composites; a method of presentation that could otherwise produce a range effect (e.g., Poulton, 1975).

Across experiments, the likeness rating scale varied: in Experiment 1, ratings were provided on a 15-point scale anchored from '*very dissimilar'* to '*very alike'*, while in subsequent

experiments, a truncated scale, with better-defined endpoints, was used (i.e., (1 = *very poor likeness* … 7 = *very good likeness*). This decision arose as Experiment 1's data revealed unequal distribution of ratings across the scale, with participant's evidencing reluctance to rate with higher scale points (from 8 – 15). For the ensuing GEE and GLMM analyses, this necessitated scale-recoding; specifically scale points of 8 and above were collapsed to a single category (scale point 8) to produce a more equal frequency distribution across the remaining scale points. We hoped to avoid scale recoding in subsequent experiments, as this action reduces the range and veracity of the data. However, participants in Experiments 2 – 3 still infrequently selected the highest scale point (of 7) and so similar, although less extreme, value-collapsing was undertaken (i.e., in Experiment 2, scale ratings from 5 – 7 were collapsed to a value of 5; and in Experiment 3, scale ratings of 6 and 7 were collapsed to a value of 6). Dependent on condition assignment, participants in Experiment 4 demonstrated a reluctance to use lower and higher scale points, respectively, thus for all participant responses values from 1 to 3 were recoded as 3, and values 5 to 7 as 5.

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

4.1 *Power and Inferential Analyses*

4.1.1 Naming Analyses

Approach

Generalized Estimating Equations (GEE) were used to analyse participant naming responses to composites for all experiments presented in this paper (SPSS Version 29 using GENLIN, IBM Corp.). This regression technique uses a binary approach to composite naming responses. Two main analyses were conducted, one for *correct* naming (coded as *1* when the given name was accurate, and *0* otherwise) and the other for *mistaken* naming (coded as *1*

when the given name was erroneous, and *0* otherwise), with a consideration of both indices affording a comprehensive assessment of composite quality.

For all experiments, two GEE analyses were first conducted by the second author and checked by the last. The first analysis was by-*participants*, a conventional analysis to assess the extent to which results generalise to other participants. The second, by-*items,* to confirm that results generalise to other stimuli, thus avoiding suggestion of a stimuli-as-a-fixed-effect fallacy (Clark, 1973). These analyses were modelled by specifying the coding for the within-subjects' variable as *items* (identities or stimuli in the experiment) in the former, and *participant-namers* in the latter. Both analyses produced the same pattern of significant and non-significant differences, except for one additional significant difference for (the less forensically-important) mistaken naming measure in the by-items analysis in Experiment 2, and so, for brevity, by-participant analyses are presented in Results, with further details provided in Appendix A, and by-items analyses in Appendix B.

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

In each analysis, similar to Repeated Measures ANOVA, participant responses were modelled as being equally correlated, achieved by selecting an Exchangeable Working Correlation Matrix. Unlike tests such as ANOVA, regression models are usually subject to an iterative process to select predictors. As such, to lessen the chance of making a Type II error, predictors (IVs) were maintained in the model based on the established criteria for regression analyses of $p < .1$ (e.g., Field, 2018). Both Model-based and Robust covariance estimators were conducted, with smaller standard error (*SE)* values for a predictor's coefficient (*B)* indicating a better overall fit of the data. *SE(B)* values emerged much lower for Model-based (cf. Robust), or varied little, and so, as Model-based is available in more statistical packages, this estimator was selected throughout. Further, for all analyses, coefficients, standard errors and confidence intervals were checked for appropriate values, neither too low nor too high, that might otherwise indicate an issue with model fit.

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

Using the above approach, we also took this opportunity to conduct analyses using GLMM (Appendices C-E). This regression approach involves fixed effects (predictors, or IVs, as modelled by GEE), but also random effects (e.g., the influence of participants and stimuli). As such, it provides a combined by-participants and by-items model that is gaining popularity (Meteyard & Davies, 2020). At the time of writing, GLMM only seems to have been used to formally analyse responses to composites in one prior publication (Erickson et al., 2022), and

so we compared the established GEE method with GLMM to provide evidence for or against the applicability of the latter technique.

Statistical Power

A between-subjects design was followed for face construction (Stage 1) and composite naming (Stage 2), with appropriate Generalized Estimating Equations (GEE) analyses planned. To be of practical significance, at least a medium effect size was desired. Previous, similar work (e.g., Erickson et al., 2022; Frowd et al., 2013; Martin et al., 2017; Portch et al., 2017; Skelton et al., 2020) indicated that a minimum of 10 participants per condition was required for face construction and composite naming, respectively, with the appropriateness of these estimates assessed by computer simulation.

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

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

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

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

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

Baseline performance was defined relative to immediate construction for an expected mean correct naming of 30% for a computerised feature system (Frowd et al., 2015). It was realised for the model's Constant (B_0) by random sampling of a Normal distribution based on a value of -0.85, with *SD* set to 0.1 to give a sensible range (+/- 2 *SD*) from 26 to 34% between participant-namers. Based on expectation, *Exp(B)* was modelled to *reduce* naming successively by a medium effect across each delay interval (sampling *B* from a random Normal distribution with mean values of -0.92, -1.83 and -2.75, respectively), again with *SD* = 0.1, to provide variability in participant-namer responses. Residual errors (e_{ij}) were added to each participant-namer response, again using a random Normal distribution ($M = 0.0$), $SD = 0.5$, again to provide suitably variable individual responses. Finally, as target identities are sometimes not correctly named (typically 1 in 20), we modelled this situation, since associated composite responses cannot be correct and so are removed prior to analyses—a procedure that increases *SE(B)* and impacts statistical power. Accordingly, 5% of cases were selected by chance to be an unfamiliar identity and then processed accordingly. Simulation included three random effects: stimulus items (coded 1-10), participant-witnesses (1-40), and participant-namers (1-40).

Retention Interval was significant as a main effect (i.e., with an omnibus value of $p < .05$) for each simulation, by-participants and by-items. Reverse Helmert contrasts emerged significant

the vast majority of the time, 91% by-participants and 94% by-items. Power was weakest for the first contrast (i.e., 3-4 hr vs. immediate) and was significant 76% of the time by-participants and 84% by-items; other contrasts were significant over 99%. These simulations indicate that good statistical power has been achieved.

Experiment 2 involved a factorial design with predictors of *Early Recall* and *Interview Type* (*see* Equation 2, below). Computer simulation was based on a medium, positive, additive effect for these two predictors (i.e., $Exp(B) = +2.5$) using the proposed design (e.g., 10 different stimuli items, and 10 participants / group for both participant-witnesses and participant-namers). Baseline performance for PRO-fit was taken from Experiment 1 at the two-day delay interval, a mean of 9% correct (i.e., $B_0 = -2.31$); other parameters were the same as described for Experiment 1, above (e.g., same settings for SD). Simulation by-participants and by-items revealed that these two predictors were significant between 95 and 97% of the time, again indicating good statistical power.

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

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

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

Experiment 3 involved a single factor, *Early Recall*. Relative to computerised feature systems, composites from Sketch are usually constructed more effectively at a long retention interval (e.g., M = ~15% in Frowd et al., 2015, and ~35 - 45% in Kuivaniemi-Smith, 2023), and so a medial baseline of 30% correct was specified, giving $B_0 = -0.85$. Using other settings from the first simulation and modelling a medium effect for the predictor, *Early Recall*, this

fixed effect was significant 83% by-participants and 84% by-items, once again indicating good statistical power.

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

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

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

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

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

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

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

So, overall, while the estimated sample sizes may seem small, they have been successfully used in previous research (e.g., see references above), and are here supported by simulation.

Indeed, this sample size was able to reliably detect a medium effect in each of the experiments reported in this paper (see General Discussion and Appendix D). Also, it was sufficient for analysing correct naming responses using a complementary regression technique, Generalized Linear Mixed Models (GLMM; see Appendix C).

4.1.2 Likeness Ratings

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

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

References

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127 –135.

Brown, C., Portch, E., Nelson, L., & Frowd, C. D. (2020). Reevaluating the role of verbalization of faces for composite production: Descriptions of offenders matter! *Journal of Experimental Psychology: Applied, 26*, 248–265.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Erickson, W. B., Brown, C., Portch E., Lampinen, J. M., Marsh, J. E., Fodarella, C., Petkovic, A., Coultas, C., Newby, A., Date, L., Hancock, P. J. B., & Frowd, C. D. (2022). The impact of weapons and unusual objects on the construction of facial composites, *Psychology, Crime & Law*, *30*(3), 207–228.

Field, A. (2018). *Discovering statistics using SPSS*. 5th Ed. Sage: London.

Fodarella, C., Frowd, C. D., Warwick, K., Hepton, G., Stone, K., Date, L., & Heard, P. (2017). Adjusting the focus of attention: helping witnesses to evolve a more identifiable composite. *Forensic Research & Criminology International*, *5*(1), 00143

Fodarella, C., Kuivaniemi-Smith, H. J., Gawrylowicz, J., & Frowd, C. D. (2015). Forensic procedures for facial-composite construction. *Journal of Forensic Practice, 17*, 259–270.

Fodarella, C., Marsh, J. E., Chu, S., Athwal-Kooner, P., Jones, H. S., Skelton, F. C., Wood, E. Jackson, E., & Frowd, C. D. (2021). The importance of detailed context reinstatement for the production of identifiable composite faces from memory. *Visual Cognition*, *29*(3), 180–200.

Frowd, C. D. (2021). Forensic Facial Composites. In Toglia, M., Smith, A., & Lampinen, J. M. (Eds.) *Methods, Measures, and Theories in Forensic Facial-Recognition* (pp. 34–64). Taylor and Francis: UK.

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

Frowd, C. D., Erickson, W. B., Lampinen, J. L., Skelton, F. C., McIntyre, A. H., & Hancock, P. J. B. (2015). A decade of evolving composite techniques: regression- and meta-analysis. *Journal of Forensic Practice*, *17*, 319–334.

Frowd, C. D., Nelson, L., Skelton F. C., Noyce, R., Atkins, R., Heard, P., Morgan, D., Fields, S., Henry, J., McIntyre, A., & Hancock, P. J. B. (2012a). Interviewing techniques for Darwinian facial composite systems. *Applied Cognitive Psychology*, *26*, 576–584.

Frowd, C. D., Skelton, F. C., Hepton, G., Holden, L., Minahil, S., Pitchford, M., McIntrye, A., Brown, C., & Hancock, P. J. B. (2013). Whole-face procedures for recovering facial images from memory. *Science & Justice*, *53*(2), 89–97.

IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 29.0. Armonk, NY: IBM Corp

Kuivaniemi-Smith, H. J. (2023). Understanding and improving the effectiveness of sketch facial composites. [PhD Thesis]. University of Lancashire, UK.

Kuivaniemi-Smith, H. J., Nash, R. A., Brodie, E. R., Mahoney, G., & Rynn, C. (2014). Producing facial composite sketches in remote cognitive interviews: A preliminary investigation. *Psychology, Crime & Law*, *20*, 389–406.

Martin, A. J., Hancock, P. J. B., & Frowd, C. D. (2017). Breath, relax and remember: an investigation into how focused breathing can improve identification of EvoFIT facial composites. *In Proceedings of the 2017 Seventh International Conference on Emerging Security Technologies (EST)* (pp. 79–84). Institute of Electrical and Electronics Engineers.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Portch, E., Logan, K., & Frowd, C. D. (2017). Interviewing and visualisation techniques: attempting to further improve EvoFIT facial composites. In *Proceedings of the 2017 Seventh International Conference on Emerging Security Technologies (EST)* (pp. 97–102). Institute of Electrical and Electronics Engineers.

Poulton, E. C. (1975). Range effects in experiments on people. *American Journal of Psychology*, *88*, 3–32.

Richardson, B. H., Brown, C., Heard, P., Pitchford, M., Portch, E., Lander, K., Marsh, J. E., Bell, R., Fodarella, C., Taylor, S. A., Worthington, M., Ellison, L., Charters, P., Green, D., Minahil, S., & Frowd, C. D. (2020). The advantage of low and medium attractiveness for facial composite production from modern forensic systems. *Journal of Applied Research in Memory and Cognition*, *9*(3), 381–395.

Sauerland, M., Holub, F., & Sporer, S. (2008). Person descriptions and person identifications: Verbal overshadowing or recognition criterion shift? *European Journal of Cognitive Psychology*, *20*, 497–528.

Skelton, F. C., Frowd, C. D., Hancock, P. J. B., Jones, H. S., Jones, B. C., Fodarella, C., Battersby, K., & Logan, K. (2020). Constructing identifiable composite faces: The

importance of cognitive alignment of interview and construction procedure. *Journal of Experimental Psychology: Applied, 26*, 507–521.

Turtle, J. W., & Yuille, J. C. (1994). Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hypermnesia. *Journal of Applied Psychology*, *79*(2), 260–271.