Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: www.tandfonline.com/journals/terg20

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

Emma Portch, Charity Brown, Cristina Fodarella, Elizabeth Jackson, Peter J. B. Hancock, Colin G. Tredoux, Michael B. Lewis, Chang Hong Liu, John E. Marsh, William Blake Erickson, Nicholas Philip Mitchell, Chiara Fasching, Linda Tran, Ellena Wood, Elaine A. Damin, Leonie Robertshaw, James Michael Lampinen, Louisa Date, Spike Joyce, Leonie Brooks, Ariell Farrow, Tom Barnes & Charlie D. Frowd

To cite this article: Emma Portch, Charity Brown, Cristina Fodarella, Elizabeth Jackson, Peter J. B. Hancock, Colin G. Tredoux, Michael B. Lewis, Chang Hong Liu, John E. Marsh, William Blake Erickson, Nicholas Philip Mitchell, Chiara Fasching, Linda Tran, Ellena Wood, Elaine A. Damin, Leonie Robertshaw, James Michael Lampinen, Louisa Date, Spike Joyce, Leonie Brooks, Ariell Farrow, Tom Barnes & Charlie D. Frowd (08 Jul 2025): The impact of forensic delay: facilitating facial composite construction using an early-recall retrieval technique, Ergonomics, DOI: 10.1080/00140139.2025.2519876

To link to this article: https://doi.org/10.1080/00140139.2025.2519876

Published online: 08 Jul 2025.

Article views: 164

View Crossmark data

View supplementary material

Submit your article to this journal

View related articles

ERGONOMICS
https://doi.org/10.1080/00140139.2025.2519876

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

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

Emma Portch[a], Charity Brown[b], Cristina Fodarella[c], Elizabeth Jackson[c], Peter J. B. Hancock[d], Colin G. Tredoux[e], Michael B. Lewis[f], Chang Hong Liu[a], John E. Marsh[c,g], William Blake Erickson[h], Nicholas Philip Mitchell[i], Chiara Fasching[i], Linda Tran[c], Ellena Wood[c], Elaine A. Damin[c], Leonie Robertshaw[j], James Michael Lampinen[k], Louisa Date[c], Spike Joyce[c], Leonie Brooks[c], Ariell Farrow[c], Tom Barnes[c] and Charlie D. Frowd[c]

[a]Department of Psychology, Bournemouth University, Poole, United Kingdom of Great Britain and Northern Ireland; [b]School of Psychology, University of Leeds, Leeds, United Kingdom of Great Britain and Northern Ireland; [c]School of Psychology, University of Lancashire, Preston, United Kingdom of Great Britain and Northern Ireland; [d]Department of Psychology, Faculty of Natural Sciences, University of Stirling, Stirling, United Kingdom of Great Britain and Northern Ireland; [e]Department of Psychology, University of Cape Town, Rondebosch, South Africa; [f]School of Psychology, Cardiff University, Cardiff, United Kingdom of Great Britain and Northern Ireland; [g]Department of Health, Learning and Technology, Luleå University of Technology, Luleå, Sweden; [h]Department of Life Sciences, Texas A&M University, San Antonio, TX, USA; [i]School of Engineering and Computing, University of Lancashire, Preston, United Kingdom of Great Britain and Northern Ireland; [j]School of Life Sciences, University of Dundee, Dundee, United Kingdom of Great Britain and Northern Ireland; [k]Psychological Science University of Arkansas, Fayetteville, AR, USA

ABSTRACT

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 retention interval impacts 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.

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.

ARTICLE HISTORY
Received 9 February 2024
Accepted 6 June 2025

KEYWORDS
Facial composite; face memory; retention interval; self-administered interview; holistic interview

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;

CONTACT Emma Portch eportch@bournemouth.ac.uk Department of Psychology, Bournemouth University, Poole BH12 5BB, United Kingdom of Great Britain and Northern Ireland

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00140139.2025.2519876.

Solomon and 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 and 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, McQuiston-Surrett, et al. 2005; Frowd, Nelson, et al. 2012). 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, Nelson, et al. 2012).

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 and Rienick 1998; Ellis, Shepherd, and Davies 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, Carson, Ness, Richardson, et al. 2005, Frowd, McQuiston-Surrett, et al. 2005; Frowd et al. 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, Frowd, and Portch (2017) found that PRO-fit composites were better recognised when participant-witnesses verbally described the face to a practitioner during two 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 and Karpicke 2006), a concept inherent within the Transfer Appropriate Processing (TAP) framework (Morris, Bransford, and Franks 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, Trevisan, and Sundararajan 2017; Yang et al. 2021). Applying this theory to the Brown, Frowd, and Portch (2017) findings, elicitation of early recall first likely prevents decay of face memory (Ellis, Shepherd, and Davies 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 and Karpicke 2006), despite their progressively increased temporal displacement from face encoding (Whitten and 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, McQuiston-Surrett, et al. 2005; Frowd, Nelson, et al. 2012).

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, Hope, and Fisher 2009), before marked face-memory decay (Ellis, Shepherd, and Davies 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 and Engstler-Schooler 1990) – description accuracy may be an important mediator of such effects (Meissner, Brigham, and Kelley 2001). By this account, feature-based recall is thought to overwrite the originally-encoded visual face memory trace, thus highly-accurate verbal templates may aid composite construction (e.g. Meissner, Brigham, and Kelley 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 and 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 and 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, Bransford, and Franks 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 and 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. extraversion). 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 and Gibson 2013), it is of practical importance to assess whether it remains useful 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 and additive 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 hypothesised 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, 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 their possible influence: 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 (Table 1), with further detail available in interim method sections and online Supplementary Materials.

Experiment 1: PRO-fit face construction by retention interval

Introduction

Experiment 1 compared composite effectiveness when face construction occurred immediately, and at three forensically-typical retention intervals: 3-4 hours was chosen as the shortest likely interval (Frowd, Carson, Ness, Richardson, et al. 2005); two days was chosen as more typical (Frowd, McQuiston-Surrett, et al. 2005); and one week was also used, as this interval often occurs in

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 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.

serious criminal cases (Frowd, Pitchford, et al. 2012). 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.

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.

Correct Naming: Generalised 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, Skelton, et al. 2012), 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 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* ≥ 0.43, *MD* = −0.03 to 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

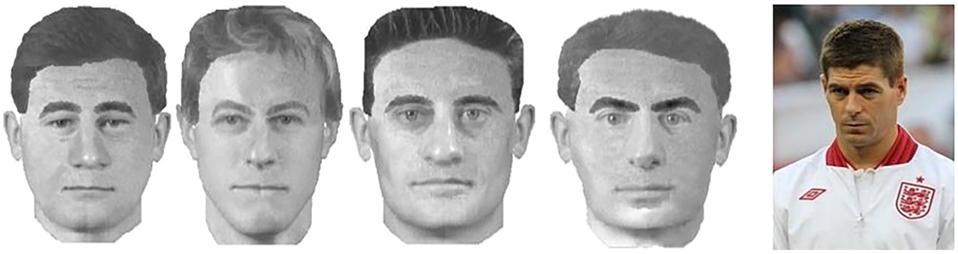

Figure 1. Example composites constructed to resemble the UK footballer Steven Gerrard. Each composite was created by a different person after experiencing one of four post-encoding retention intervals, from left-to-right: immediate, 3-4 hours, 2 days and 1 week. For reasons of copyright, the actual target 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 presented a more frontal view of the face).

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 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, Shepherd, and Davies 1980), reducing composite effectiveness (Frowd et al. 2015).

Table 2. Experiment 1 results for each DV (correct naming, mistaken naming and likeness ratings) by increasing *Retention Interval*.

	Retention interval[a]			
DV	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 a percentage, and (in parentheses) as a number of correct names offered (numerator) out of the number of correctly identified targets (denominator). $^a p < .001$. Mistaken Naming: Shown as a percentage, and (in parentheses) as a number of mistaken names offered (numerator) out of the number of correctly identified targets (denominator). $^a p < .02$. Likeness Ratings: Rating scale (1 = very dissimilar … 15 = very alike; with scale points 9–15 recoded as 8). Shown are mean likeness ratings and (in parentheses) SE. $^a p < .01$. 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, where an ascending sorting order was used); 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.

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 2, *Note*), techniques that mitigate the documented sharp decline in memory that occurs between these two timepoints (e.g. Ellis, Shepherd, and Davies 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, Shepherd, and Davies 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, McQuiston-Surrett, et al. 2005; Frowd, Nelson, et al. 2012). 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).

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 = 0.1][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.

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

Table 3. Percentage correct naming of PRO-fit composites constructed by *Early Recall* and *Interview Type*.

	Early recall[a]		
Interview type[b]	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. [a,b]$p < .001$.

Table 4. Model parameters for the impact of *Early Recall* and *Interview Type* on correct naming of PRO-fit composites.

Fixed effects	*B*	*SE(B)*	χ_1^2 *(1)*	*p*	*Exp(B)*	95% CI(−)	95% CI(+)
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) are presented by 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 and 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.

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).

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 (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.

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* (*p*s < .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.

Table 5. Percentage mistaken naming of PRO-fit composites by *Early Recall and Interview Type*.

	Early recall		
Interview type^a	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]. $^a p < .05$.

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* ≤ .005) except $^a p = 1.0$.

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, Frowd, and Portch (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 and Little 1990; Laughery, Duval, and Wogalter 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 identity and age information 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.

Table 7. Percentage correct naming of sketch composites constructed by *Early Recall*.

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

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

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].

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 8. Model parameters for the impact of *Early Recall* on correct naming for sketch composites.

Fixed effects	*B*	*SE(B)*	$\chi_1^2(1)$	*p*	*Exp(B)*	95% CI(−)	95% CI(+)
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[a]	
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)].
[a] *p* < .001.

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, Nelson, et al. 2012). 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, Frowd, and Portch 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 and Lloyd-Jones 2002; Frowd and Fields 2011; MacLin 2002; Schooler and 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) [$\chi_1^2(2)=80.03$, $p<.001$] (Table 10). Parameter estimates (Table 11) revealed differences between all three conditions ($ps<.001$), with ER-H-CI performing best, followed by H-CI, and CI performing worst.

Table 10. Percentage correct naming of EvoFIT composites constructed by *Interview Type*.

Interview type^{a}		
CI	H-CI	ER-H-CI
28.9 (24/83)	45.5 (40/88)	71.6 (63/88)

Note. $^{a}p<.001$: all comparisons.

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* [$\chi_1^2(2)=7.476$, $p=.024$]. 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 ($p=.85$).

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).

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

Table 12. Percentage mistaken naming of EvoFIT composites constructed by *Interview Type*.

Interview type^{a}		
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 [$B=-1.68$, $SE(B)=0.28$].
$^{a}p<.05$.

Table 13. Mean likeness ratings (*SE*) of EvoFIT composites constructed by *Interview Type*.

Interview type^{a}		
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)].
$^{a}p<.001$: All comparisons.

Table 11. Model parameters for the impact of *Interview Type* on correct naming of EvoFIT composites.

Fixed effects	*B*	*SE(B)*	$\chi_1^2(1)$	*p*	*Exp(B)*	95% CI(−)	95% CI(+)
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 [$B=-0.89$, $SE(B)=0.15$].

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.

The statistical approach remained consistent with previous analyses, incorporating data from Experiments 2–4 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).

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; ER-H-CI: early recall + face and holistic recall. Values are expressed in percentages for Correct Naming and Mistaken Naming, while mean ratings are presented for Likeness Rating.

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 size [$B = -0.84$, $SE(B) = 0.15$, $1/Exp(B) = 1.38$ (1.00, 1.90)]. This IV was not retained in the model by-items.

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, Shepherd, and Davies 1980). While Brown, Frowd, and Portch (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 and Matsuo 2021), perhaps producing a recoded verbal template that less effectively guides composite construction (Meissner, Brigham, and Kelley 2001)[6]. 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, Nelson, et al. 2012; 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.

In contrast, 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; Frowd, Nelson, et al. 2012). 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 and Little 1990; Laughery, Duval, and Wogalter 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, Bruce, and Johnston 1990; Valentine 1991).

The system-wide benefit of early recall likely does not solely reflect early enhancement of face memory. If it did, a face-recall Cognitive Interview conducted soon after encoding should facilitate subsequent construction regardless of later recall attempts. However, Brown, Frowd, and Portch (2017) found no such facilitation when participant-witnesses recalled the face only *once*, at a 3-4hour 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 Transfer Appropriate Processing (TAP; e.g. Adesope, Trevisan, and Sundararajan 2017; Roediger and 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 and Wolters 2006; Whitten and 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 an eye-region focus interview was adopted [see Skelton et al, 2020; 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 TAP principles, pre-construction holistic recall may specifically prime recognition, rather than recall, mechanisms; the two often considered separable (e.g. Wells and Hryciw 1984). For feature systems, primed recognition may help witnesses to assess when the created image resembles the target (e.g. Frowd et al. 2008, 2015; Frowd, Nelson, et al. 2012). 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, Nelson, et al. 2012), but impedes earlier ones, specifically, selection of whole-faces from arrays that best resemble the target (e.g. Brown and Lloyd-Jones 2002; Frowd and 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 and 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. university 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 and Yuille 1994) thought about the face and/or reviewed/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 Generalised Linear Mixed-Effects Models (GLMM; 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. Indeed, by simultaneously adopting a by-participants and by-items approach, they circumvent the difficulties of attempting to reconcile these trends when they are disparate. 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.

Notes

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 $\alpha \leq 0.1$, while $\alpha \leq 0.05$ was used for subsequent post-hoc and simple-main effect tests (e.g. Field, 2018). Also, see Appendix A for more infor-

mation regarding the by-participants analyses and Appendix B for complementary analyses by-items. Appendix C presents analyses instead using Generalised Linear Mixed-Effects Models (GLMM).

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.
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 (Appendix E).
4. When an interaction or an IV is removed from a model, this indicates that the variable does not hold explanatory value for 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).
6. 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].
7. We acknowledge that the results would have been better presented as a series of tables; however, this format was not possible due to publication constraints.
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.

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.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported that there is no funding associated with the work featured in this article.

ORCID

Emma Portch http://orcid.org/0009-0008-7895-8100
Charity Brown http://orcid.org/0000-0001-9697-4878
Cristina Fodarella http://orcid.org/0000-0001-5551-3450
Peter J. B. Hancock http://orcid.org/0000-0001-6025-7068
Colin G. Tredoux http://orcid.org/0000-0002-9653-786X
Michael B. Lewis http://orcid.org/0000-0002-5735-5318
Chang Hong Liu http://orcid.org/0000-0002-2426-4014
John E. Marsh http://orcid.org/0000-0002-9494-1287
William Blake Erickson http://orcid.org/0000-0002-2765-3699
James Michael Lampinen http://orcid.org/0000-0002-5854-521X
Charlie D. Frowd http://orcid.org/0000-0002-5082-1259

Data availability statement

Raw data for these experiments were generated at the Universities of Lancashire, Leeds and Dundee. These data are available from the corresponding author [EP] on request.

References

Adesope, O. O., D. A. Trevisan, and N. Sundararajan. 2017. "Rethinking the Use of Tests: A Meta-Analysis of Practice Testing." *Review of Educational Research 87* (3): 659–701. doi:10.3102/0034654316689306.

Agresti, A., J. G. Booth, J. P. Hobert, and B. Caffo. 2000. "Random-Effects Modelling of Categorical Response Data." *Sociological Methodology 30* (1): 27–80. doi:10.1111/0081-1750.t01-1-00075.

Ashkenazi, T., and R. P. Fisher. 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* (2): 200–208. doi:10.1037/h0101871.

Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily. 2013. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal." *Journal of Memory and Language 68* (3): 255–278. doi:10.1016/j.jml.2012.11.001.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M H. H. Stevens, and J.-S. S. White. 2009. "Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution." *Trends in Ecology & Evolution 24* (3): 127–135. doi:10.1016/j.tree.2008.10.008.

Brown, C., and T. J. Lloyd-Jones. 2002. "Verbal Overshadowing in a Multiple Face Presentation Paradigm: Effects of Description Instruction." *Applied Cognitive Psychology 16* (8): 873–885. doi:10.1002/acp.919.

Brown, C., C. D. Frowd and E. Portch. 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)*, 38–43. Canterbury, UK: Institute of Electrical and Electronics Engineers. doi:10.1109/EST.2017.8090396.

Brown, C., E. Portch, L. Nelson, and C. D. Frowd. 2020. "Reevaluating the Role of Verbalization of Faces for Composite Production: Descriptions of Offenders Matter!" *Journal of Experimental Psychology. Applied 26* (2): 248–265. doi:10.1037/xap0000251.

Burton, A. M., V. Bruce, and R. A. Johnston. 1990. “Understanding Face Recognition with an Interactive Activation Model.” *British Journal of Psychology* 81 (3): 361–380. doi:10.1111/j.2044-8295.1990.tb02367.x.

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* (4): 335–359. doi:10.1016/S0022-5371(73)80014-3.

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., and M. Little. 1990. “Drawing on Memory: Exploring the Expertise of a Police Artist.” *Medicine, Science, and the Law 30* (4): 345–353. doi:10.1177/106002809003000412.

Ebbesen, E. B., and C. B. Rienick. 1998. “Retention Interval and Eyewitness Memory for Events and Personal Identifying Attributes.” *The Journal of Applied Psychology* 83 (5): 745–762. doi:10.1037/0021-9010.83.5.745.

Ellis, H. D., J. W. Shepherd, and G. M. Davies. 1980. “The Deterioration of Verbal Descriptions of Faces over Different Delay Intervals.” *Journal of Police Science and Administration 8*: 101–106.

Erickson, W. B., C. Brown, E. Portch, J. M. Lampinen, J. E. Marsh, C. Fodarella, A. Petkovic, C. Coultas, A. Newby, L. Date, P. J. B. Hancock, and C. D. Frowd. 2024. “The Impact of Weapons and Unusual Objects on the Construction of Facial Composites.” *Psychology, Crime & Law 30* (3): 207–228. doi:10.1080/1068316X.2022.2079643.

Field, A. 2018. *Discovering Statistics Using SPSS*. 5th ed. London: Sage.

Fisher, R. P., R. E. Geiselman, D. S. Raymond, L. M. Jurkevich, and M. L. Warhaftig. 1987. “Enhancing Enhanced Eyewitness Memory: Refining the Cognitive Interview.” *Journal of Police Science and Administration 15*: 291–297.

Fodarella, C., C. D. Frowd, K. Warwick, G. Hepton, K. Stone, L. Date, and P. Heard. 2017. “Adjusting the Focus of Attention: Helping Witnesses to Evolve a More Identifiable Composite.” *Forensic Research & Criminology International 5* (1): 00143.

Fodarella, C., J. E. Marsh, S. Chu, P. Athwal-Kooner, H. S. Jones, F. C. Skelton, E. Wood, E. Jackson, and C. D. Frowd. 2021. “The Importance of Detailed Context Reinstatement for the Production of Identifiable Composite Faces from Memory.” *Visual Cognition 29* (3): 180–200. doi:10.1080/13506285.2021.1890292.

Frowd, C. D. 2021. “Forensic Facial Composites.” In *Methods, Measures, and Theories in Forensic Facial-Recognition*, edited by A. M. Smith, M. P. Toglia, and J. M. Lampinen, 34–64. Oxfordshire, UK: Taylor and Francis.

Frowd, C. D. 2023. “Eyewitnesses and the Use and Application of Cognitive Theory.” In *Introduction to Applied Psychology*, edited by G. Davey, 207–232. Chichester, UK:BPS Wiley-Blackwell.

Frowd, C. D., V. Bruce, H. Ness, L. Bowie, C. Thomson-Bogner, J. Paterson, A. McIntyre, and P. J. B. Hancock. 2007. “Parallel Approaches to Composite Production.” *Ergonomics 50* (4): 562–585. doi:10.1080/00140130601154855.

Frowd, C. D., V. Bruce, A. Smith, and P. J. B. Hancock. 2008. “Improving the Quality of Facial Composites Using a Holistic Cognitive Interview.” *Journal of Experimental Psychology. Applied 14* (3): 276–287. doi:10.1037/1076-898X.14.3.276.

Frowd, C. D., D. Carson, H. Ness, D. McQuiston, J. Richardson, H. Baldwin, and P. J. B. Hancock. 2005. “Contemporary Composite Techniques: The Impact of a Forensically-Relevant Target Delay.” *Legal and Criminological Psychology 10* (1): 63–81. doi:10.1348/135532504X15358.

Frowd, C. D., D. Carson, H. Ness, J. Richardson, L. Morrison, S. McLanaghan, and P. J. B. Hancock. 2005. “A Forensically Valid Comparison of Facial Composite Systems.” *Psychology, Crime & Law 11* (1): 33–52. doi:10.1080/10683160310001634313.

Frowd, C. D., W. B. Erickson, J. L. Lampinen, F. C. Skelton, A. H. McIntyre, and P. J. B. Hancock. 2015. “A Decade of Evolving Composite Techniques: Regression- and Meta-Analysis.” *Journal of Forensic Practice 17* (4): 319–334. doi:10.1108/JFP-08-2014-0025.

Frowd, C. D., and S. Fields. 2011. “Verbalisation Effects in Facial Composite Production.” *Psychology, Crime & Law 17* (8): 731–744. doi:10.1080/10683161003623264.

Frowd, C. D., D. McQuiston-Surrett, I. Kirkland, and P. J. B. Hancock. 2005. “The Process of Facial Composite Production.” In *Forensic Psychology and Law*, edited by A. Czerederecka, T. Jaskiewicz-Obydzinska, R. Roesch, and J. Wojcikiewicz, 140–152. Krakow: Institute of Forensic Research Publishers.

Frowd, C. D., L. Nelson, F. C. Skelton, R. Noyce, R. Atkins, P. Heard, D. Morgan, S. Fields, J. Henry, A. McIntyre, and P. J. B. Hancock. 2012. “Interviewing Techniques for Darwinian Facial Composite Systems.” *Applied Cognitive Psychology 26* (4): 576–584. doi:10.1002/acp.2829.

Frowd, C. D., M. Pitchford, V. Bruce, S. Jackson, G. Hepton, M. Greenall, A. McIntyre, and P. J. B. Hancock. 2011. “The Psychology of Face Construction: Giving Evolution a Helping Hand.” *Applied Cognitive Psychology 25* (2): 195–203. doi:10.1002/acp.1662.

Frowd, C. D., M. Pitchford, F. C. Skelton, A. Petkovic, C. Prosser, and B. Coates. 2012. “Catching Even More Offenders with EvoFIT Facial Composites.” In *Proceedings of the 2012 Third International Conference on Emerging Security Technologies (EST)*, 20–26. Lisbon, Portugal: Institute of Electrical and Electronics Engineers. doi: 10.1109/EST.2012.26.

Frowd, C. D., E. Portch, A. Killeen, L. Mullen, A. J. Martin, and P. J. B. Hancock. 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)*, 1–7. Colchester, UK: Institute of Electrical and Electronics Engineers. doi:10.1109/EST.2019.8806211.

Frowd, C. D., F. C. Skelton, C. Atherton, M. Pitchford, G. Hepton, L. Holden, A. McIntyre, and P. J. B. Hancock. 2012. “Recovering Faces from Memory: The Distracting Influence of External Facial Features.” *Journal of Experimental Psychology-Applied 18* (2): 224–238. doi:10.1037/a0027393.

Frowd, C. D., F. C. Skelton, G. Hepton, L. Holden, S. Minahil, M. Pitchford, A. McIntyre, C. Brown, and P. J. B. Hancock. 2013. “Whole-Face Procedures for Recovering Facial Images from Memory.” *Science & Justice: Journal of the Forensic Science Society 53* (2): 89–97. doi:10.1016/j.scijus.2012.12.004.

Gabbert, F., L. Hope, and R. Fisher. 2009. “Protecting Eyewitness Evidence: Examining the Efficacy of a Self-Administered Interview.” *Law and Human Behavior 33* (4): 298–307. doi:10.1007/s10979-008-9146-8.

Gill, J., and G. King. 2004. “What to Do When Your Hessian is Not Invertible.” *Sociological Methods & Research 33* (1): 54–87. doi:10.1177/0049124103262681.

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]*. New York: IBM Corp.

Kellogg, R. T. 2007. “Are Written and Spoken Recall of Text Equivalent?” *The American Journal of Psychology 120* (3): 415–428. doi:10.2307/20445412.

Kuivaniemi-Smith, H. J. 2023. “Understanding and Improving the Effectiveness of Sketch Facial Composites”. Doctoral Diss., University of Lancashire.

Laughery, K. R., C. Duval, and M. S. Wogalter. 1986. “Dynamics of Facial Recall.” In *Aspects of Face Processing*, edited by H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, 373–387. Dordrecht, The Netherlands: Martinus Nijhoff.

MacLin, M. K. 2002. “The Effects of Exemplar and Prototype Descriptors on Verbal Overshadowing.” *Applied Cognitive Psychology 16* (8): 929–936. doi:10.1002/acp.923.

Meissner, C. A., J. C. Brigham, and C. M. Kelley. 2001. “The Influence of Retrieval Processes in Verbal Overshadowing.” *Memory & Cognition 29* (1): 176–186. doi:10.3758/bf03195751.

Meteyard, L., and R. A. I. Davies. 2020. “Best Practice Guidance for Linear Mixed-Effects Models in Psychological Science.” *Journal of Memory and Language 112*: 104092. doi:10.1016/j.jml.2020.104092.

Miura, H., and K. Matsuo. 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. doi:10.1002/acp.3856.

Morris, C. D., J. D. Bransford, and J. J. Franks. 1977. “Levels of Processing versus Transfer Appropriate Processing.” *Journal of Verbal Learning and Verbal Behavior 16* (5): 519–533. doi:10.1016/S0022-5371(77)80016-9.

Odinot, G., A. Memon, D. La Rooy, and A. Millen. 2013. “Are Two Interviews Better than One? Eyewitness Memory across Repeated Cognitive Interviews.” *PLOS One 8* (10): e76305. doi:10.1371/journal.pone.0076305.

Odinot, G., and G. Wolters. 2006. “Repeated Recall, Retention Interval and the Accuracy-Confidence Relation in Eyewitness Memory.” *Applied Cognitive Psychology 20* (7): 973–985. doi:10.1002/acp.1263.

Osborne, J. W. 2016. *Regression & Linear Modeling: Best Practices and Modern Methods*. Los Angeles, CA: Sage Publications.

Pitchford, M., D. Green, and C. D. Frowd. 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)*, 185–190. Canterbury, UK: Institute of Electrical and Electronics Engineers. doi:10.1109/EST.2017.8090421.

Reed, P., and Y. Wu. 2013. “Logistic Regression for Risk Factor Modelling in Stuttering Research.” *Journal of Fluency Disorders 38* (2): 88–101. doi:10.1016/j.jfludis.2012.09.003.

Roediger, H. L., III, and J. D. Karpicke. 2006. “Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention.” *Psychological Science 17* (3): 249–255. doi:10.1111/j.1467-9280.2006.01693.x.

Sauerland, M., A. C. Krix, N. van Kan, S. Glunz, and A. Sak. 2014. “Speaking is Silver, Writing is Golden? The Role of Cognitive and Social Factors in Written versus Spoken Witness Accounts.” *Memory & Cognition 42* (6): 978–992. doi:10.3758/s13421-014-0401-6.

Sauerland, M., and S. L. Sporer. 2011. “Written vs. Spoken Eyewitness Accounts: Does Modality of Testing Matter?” *Behavioral Sciences & the Law 29* (6): 846–857. doi:10.1002/bsl.1013.

Schooler, J. W., and T. Y. Engstler-Schooler. 1990. “Verbal Overshadowing of Visual Memories: Some Things Are Better Left Unsaid.” *Cognitive Psychology 22* (1): 36–71. doi:10.1016/0010-0285(90)90003-m.

Shapiro, P. N., and S. D. Penrod. 1986. “Meta-Analysis of Facial Identification Rates.” *Psychological Bulletin 100* (2): 139–156. doi:10.1037/0033-2909.100.2.139.

Skelton, F. C., C. D. Frowd, P. J. B. Hancock, H. S. Jones, B. C. Jones, C. Fodarella, K. Battersby, and K. Logan. 2020. “Constructing Identifiable Composite Faces: The Importance of Cognitive Alignment of Interview and Construction Procedure.” *Journal of Experimental Psychology. Applied 26* (3): 507–521. doi:10.1037/xap0000257.

Solomon, C. J., and S. J. Gibson. 2013. “Developments in Forensic Facial Composites.” In *Advances in Forensic Human Identification*, edited by X. Mallett and T. Blythe, 235–270. Boca Raton, USA: CRC Press.

Sporer, S. L., and N. Martschuk. 2014. “The Reliability of Eyewitness Identifications by the Elderly: An Evidence-Based Review.” In *The Elderly Eyewitness in Court*, edited by M. P. Toglia, D. F. Ross, J. Pozzulo, and E. Pica, 3–37. New York: Psychology Press.

Tredoux, C.G., Y. Rosenthal, L. da Costa, and D. Nunez. 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, July 10, 1999.

Turtle, J. W., and J. C. Yuille. 1994. “Lost but Not Forgotten Details: Repeated Eyewitness Recall Leads to Reminiscence but Not Hypermnesia.” *The Journal of Applied Psychology 79* (2): 260–271. doi:10.1037/0021-9010.79.2.260.

Valentine, T. 1991. “A Unified account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition.” *The Quarterly Journal of Experimental Psychology-A, Human Experimental Psychology 43* (2): 161–204. doi:10.1080/14640749108400966.

Wells, G. L., and B. Hryciw. 1984. “Memory for Faces: Encoding and Retrieval Operations.” *Memory & Cognition 12* (4): 338–344. doi:10.3758/bf03198293.

Whitten, W. B., and R. A. Bjork. 1977. “Learning from Tests: Effects of Spacing.” *Journal of Verbal Learning and Verbal Behavior 16* (4): 465–478. doi:10.1016/S0022-5371(77)80040-6.

Yang, C., L. Luo, M. A. Vadillo, R. Yu, and D. R. Shanks. 2021. “Testing (Quizzing) Boosts Classroom Learning: A Systematic and Meta-Analytic Review.” *Psychological Bulletin 147* (4): 399–435. doi:10.1037/bul0000309.

Appendix A

Additional information for analyses using Generalised 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

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 ($QIC=920.5$, $QICC=920.5$).

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 Generalised 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[3], 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][3]. 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 (*p*s ≤ 0.007) 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* = .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

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). This 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.

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

Generalised Linear Mixed-Effects Models (GLMM)

Our approach followed the established statistical method for analysing responses to composites using GEE (e.g. Brown, Frowd, and Portch 2017; Frowd et al. 2013; Pitchford, Green, and Frowd 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. (2024) for GLMM (GENLINMIXED, SPSS Version 29, IBM. 2020, IBM 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 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. 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 when analysing multinomial responses (e.g. from ratings of likeness).

Based on a comparable design to the current experiments (Erickson et al. 2024), 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) generalise somewhat worse than separate by-participants and by-items tests using GEE (Gill and King 2004). Our expectation turned out to be true for correct naming; it was also true 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 $1-\beta > 95\%$.

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. (2024). 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
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 [$p = .016$, $SE(B) = 0.39$, $1/Exp(B) = 2.60$][3], (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 [$p = .003$, $SE(B) = 0.61$, $1/Exp(B) = 6.41$].

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* ($ps > .39$).

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[3], 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. 2024, is mentioned above.)

Experiment 2

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).

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).

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

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 (σ^2 = 0.31, *SE* = 0.26) and items (σ^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).

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 (σ^2 = 0.10, *SE* = 0.18) and items (σ^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).

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 (σ^2 = 0.79, *SE* = 0.33) and items (σ^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) = 87.46, *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 (σ^2 = 0.26, *SE* = 0.15) and items (σ^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.

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 (σ^2 = 1.45, *SE* = 0.42), items (σ^2 = 1.52, *SE* = 0.71) and experiments (σ^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 (σ^2 = 0.71, *SE* = 0.27), participant-namers (σ^2 = 0.25, *SE* = 0.15), items (σ^2 = 0.59, *SE* = 0.34) and experiments (σ^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).

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 (σ^2 = 1.81, *SE* = 0.60) and items (σ^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 (σ^2 = 0.37, *SE* = 0.25), participant-namers (σ^2 = 0.29, *SE* = 0.19) and items (σ^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.

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).

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 Generalised Linear Mixed-Effects Models (Appendices C and E). GLMM is gaining popularity in Psychology (Meteyard and Davies 2020), and has been used to analyse data from a single-experiment composite paper by Erickson et al. (2024). 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 Type* (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 (Gill and 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).

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).

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, Frowd, and Portch 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 hours 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 Early 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-hour (cf. 3-4 hour previously) retention interval, with all participant-witnesses constructing composites 48-hours 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].