Super-recognisers: face recognition performance after variable delay intervals

Josh P Davis, L. Diandra Bretfelean, Elena Belanova, and Trevor Thompson

School of Human Sciences, University of Greenwich, London, UK

Dr Josh P Davis, Reader in Applied Psychology, University of Greenwich, London, SE10 9LS, j.p.davis@gre.ac.uk; +44(0)208 331 8859

L. Diandra Bretfelean, Research Associate, University of Greenwich, London, SE10 9LS, diandra.bretfelean@outlook.com

Dr Elena Belanova, Research Associate, University of Greenwich, London, SE10 9LS, elena.belanova@greenwich.ac.uk

Dr Trevor Thompson, Associate Professor in Psychology, University of Greenwich, London, SE10 9LS, t.thompson@gre.ac.uk; +44(0)208 331 9632

Acknowledgements: This project was partly funded by the Large Scale Information Exploitation of Forensic Data (LASIE) project (*European Commission 7th Framework Programme. SEC-2013.1.6-1: 607480).* The authors would like to thank retired Detective Chief Inspector Mick Neville for organising PROMAT line-ups, and Monika Durova and Lauren Jensen, Bethan Burnside, and Nikolay Petrov for assistance with data collection. The authors would also like to thank two anonymous reviewers for their comments on an earlier draft. This project received ethical approval from the University of Greenwich Research Ethics Committee.

Conflict of interest statement: The authors declare that they have no conflict of interest.

Data Availability Statement: The Data that support the findings of this study are openly available in Open Science Framework at DOI:10.17605/OSF.IO/ZMCDH

Keywords

Policing, suspect identification, face recognition, long-term memory, super-recognition

Abstract

Outstanding long-term face recognition of suspects is a hallmark of the exceptionally skilled police 'super-recognisers' (SRs). Yet, research investigating SR's memory for faces mainly employed brief retention intervals. Therefore, in Experiment 1, 597 participants (121 SRs) viewed 10 target videos and attempted identification of targets from 10 target-present line-ups after 1 - 56 days. In Experiment 2, 1421 participants (301 SRs) viewed 20 target videos, and after a baseline of no delay to 28 days, - 10 target-present *and* 10 target-absent line-ups, to assess correct line-up rejections. Overall, delay positively correlated with hits but not with correct rejections. Most, but not all SRs, made more correct identifications and correct rejections than controls at all retention intervals, demonstrating that many SRs possess enhanced long-term face memory. This research adds to knowledge of SR's skillsets, and enhances the case for the selection of SRs to identity critical roles – particularly policing.

Super-recognisers: face recognition performance after variable delay intervals

The face recognition ability spectrum in the population ranges from *developmental prosopagnosics* who struggle to recognise familiar faces (e.g., Behrmann, Avidan, Marotta, & Kimchi, 2005; for a review see Susilo & Duchaine, 2013), to so called *super-recognisers* (SRs) with exceptional unfamiliar face processing skills (e.g., Russell et al., 2009; for a review see Noyes, Phillips, & O'Toole, 2017). Face recognition ability is inherited (e.g. Shakeshaft & Plomin, 2015; Wilmer et al., 2010), impacted by exposure (e.g., cross-age effect: Wiese, Komes, & Schweinberger, 2013; cross-ethnicity effect, Meissner & Brigham, 2001), and bears weak relationships with cognitive skills such as non-face object recognition (e.g., Bobak, Bennetts, Parris, Jansari, & Bate 2016; McCaffery, Robertson, Young, & Burton, 2018; Royer et al., 2018; Verhallen, Bosten, Goodbourn, Lawrance-Owen, Bargary, & Mollon 2017; Wilmer et al., 2010). Recent studies have evaluated SR's superiority at short-term unfamiliar face memory (e.g. Bate, Bennetts et al., 2018; Russell et al., 2009), simultaneous unfamiliar face matching (e.g. Bobak, Hancock, & Bate, 2016), and long-term *familiar* face recognition (e.g. Davis, Lander, Evans, & Jansari, 2016). However, SRs often self-report exceptional long-term *unfamiliar* face memory (e.g. Russell et al., 2009), and yet only one small-scale study has investigated this (Davis & Tamoytė, 2017). Recruiting one of the largest SR samples to date, the current research aimed to address this gap in the literature.

This research has important theoretical and applied implications. Until recently, face recognition research evaluating groups of different abilities, tended to examine prosopagnosics and/or controls scoring in the typical-ability range. Following the first empirical evaluation of SRs (Russell et al., 2009), and the understanding that the face recognition ability spectrum is far wider than previously realised, some reappraisal of theories may be necessary (i.e. those in relation to face expertise; Young & Burton, 2018).

Empirically evaluating the capabilities of those at the top end of the spectrum across different intervals of time should enhance knowledge of human face memory capacity and retention.

Practically, some police forces deploy SRs to roles utilising their superior skills. For instance, London's Metropolitan Police Service (MPS) created a full-time SR Unit. This unit is tasked with reviewing CCTV images of previously unidentified suspects, with a view to recognising these individuals or providing investigative leads. Their superior face processing abilities have been verified in empirical research (Davis et al., 2016; Davis, Treml, Forrest, & Jansari, 2018; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). When actual identity was still unknown, these officers sometimes recognised suspects in person, even in large crowds (i.e. the case of Ilhan Karatepe; O'Keefe, 2016, 22 August); or more commonly, in new footage of later crimes (i.e. the case of Austin Caballero; Robertson, 2016, April 2). In this manner, a team of seven police SRs from the MPS SR Unit made 1,071 suspect identifications in 30 months (Davis et al., 2018). For this, police SRs accessed past footage, or extracted facial stills stored on a database for simultaneous matching of those depicted. SR police working in these teams therefore needed exceptional long-term unfamiliar face recognition ability *and* simultaneous face matching ability (Robertson et al., 2016; for a review of identification from CCTV see Davis & Valentine, 2015). No data is available as to typical time between initial viewing and subsequent identification. However, the average retention interval between a crime occurring and images being available on the MPS database was about 40 days (M. Neville - former MPS Detective Chief Inspector, private communication, 2019, 6 August).

As face recognition ability is not amenable to training (e.g., Towler et al., 2019), or improved from experience in a job role requiring such skills (i.e. passport officers, White, Kemp, Jenkins, Matheson, & Burton, 2014), understanding SRs' aptitudes on empirical tasks replicating real-world deployment may assist in suitable recruitment test development (e.g.

Ramon, Bobak, & White, 2019). Finally, the descriptions of the long-term face memory abilities of the police SRs described above provides anecdotal evidence of superior skills in a small number of officers only. The current research aimed to empirically determine whether exceptional long-term face memory skills would be found in all SRs classified as such using short-term face memory tests.

Following precedent (e.g., Bobak, Bennetts et al., 2016; Bobak, Hancock et al., 2016; Russell et al., 2009), SR group inclusion criteria here were scores 2 standard deviations (SD) above the estimated population mean (i.e. top ≈ 2%) based on a representative sample (Bobak, Pampoulov, & Bate, 2016) on the short-term *Cambridge Face Memory Test: Extended* (CFMT+) (Russell et al., 2009). Participants also completed the *Glasgow Face Matching Test* (GFMT) (2010) in order to evaluate their simultaneous face matching skills, as well as to verify SR's superior face processing abilities.

Long-term face memory

Previous research has revealed correlations between short-term and long-term face memory (e.g., Bindemann, Brown, Koyas, & Russ, 2012; Davis & Tamoytė, 2017). Less research has examined face recognition accuracy using retention intervals of four weeks or more (e.g., Courtois & Mueller, 1981; Sauer, Brewer, Zweck, & Weber, 2010; Shepherd & Ellis, 1973; Shepherd, Ellis, & Davies, 1982; Yarmey, 1979; for a review see Deffenbacher, Bornstein, McGorty, & Penrod, 2008). None of these projects, however, evaluated individual differences in face recognition ability. The forgetting curve for the human face embraces a form determined by Ebbinghaus (1913) for other stimuli, although multiple factors such as interference (Deffenbacher et al., 2006), retrieval failure (Eimer et al., 2012), facial distinctiveness (Wickham, Morris, & Fritz, 2000), initial memory strength (Deffenbacher et al., 2008), and repeated exposure (e.g. Deffenbacher, Bornstein, & Penrod, 2006) can have an

impact. Most forgetting, however, occurs in the first 24 hours, and gradually increases over longer periods (Deffenbacher et al., 2008). Only one long-term face memory study has tested SRs. Davis and Tamoytė (2017) found that SRs were more accurate than controls at identification of a target from a nine-person line-up after mean retention intervals of 10 days. However, only one target-actor, sometimes in disguise, was employed as a stimulus, and participant numbers were low, limiting generalisability.

Although these results suggest that SRs' exceptional short-term memory contributes to enhanced long-term memory performance, of theoretical interest is whether the ability of SRs to retain unfamiliar face representations in memory differs from non-SRs (controls). Retention in memory, or its antithesis 'forgetting,' can be measured by participants viewing a series of faces in a learning phase and varying the interval before they are asked to identify the previously encountered faces amongst never seen faces or foils. The forgetting rate or curve can be inferred by reduced correct identification rates following each time interval.

All other things being equal, if SRs and controls display a similar forgetting curve, then SRs' long-term superiority is likely to be a direct result of their short-term memory superiority (whether it is due to enhanced encoding, enhanced retrieval, or both). As such, the level of SRs' superiority over controls after the longest retention intervals should be roughly the same as after the shortest intervals. In other words, the mean group difference between scores on a test will be similar at both time points. However, if SRs display a shallower forgetting curve, then on top of the mechanisms driving their short-term memory superiority, SRs might additionally possess more effective long-term retention skills. SR's superiority over controls after the longest retention intervals would therefore be greater than after the shortest intervals. This would be exhibited by an interaction, so that the mean difference in scores between SRs and controls would be larger after the longer interval.

The current research

The primary aim of the two experiments described here was to determine whether super-recognisers, assessed as such using a short-term face memory test and a verifying simultaneous face matching tests, would also demonstrate superior long-term face memory. Most long-term face memory research uses one target and one line-up to assess eyewitness performances. As SR unit police can encounter large numbers of unknown suspects, often displayed on moving images, participants were exposed to a series of action-varying videos of target-actors. Long-term face memory was assessed using six-person line-ups. In Experiment 1, all 10 line-ups were target-present, and planned retention intervals varied from one day to 56 days.

SRs are also more accurate than controls at correctly rejecting previously unseen faces (e.g., Bate, Bennetts et al., 2018; Davis et al. 2016). This implies that SRs may be able to better determine when stored face representations do not match with viewed faces. In policing contexts, it is important to measure the likelihood of identification by a SR, if, for instance, an innocent suspect became the target of an investigation. To measure this, Experiment 2 again employed 10 target-present line-ups. However, 10 target-absent line-ups in which the target was replaced by a foil were also included. The increase in trial numbers (n = 20) was additionally designed to better discriminate between high and low performers. Retention intervals varied from being virtually instantaneous, to provide a baseline of performance to 28 days. Identification decision confidence ratings were also analysed.

As possession of simultaneous face matching ability may be important in applied policing contexts, to measure the latter, participants completed the *Glasgow Face Matching Test* (GFMT) (Burton et al., 2010), a perceptual task with virtually no memory demands. The test's inclusion also allowed replication of previous research finding that although many SRs

are outstanding at both face memory and matching tasks, some SRs display poor simultaneous face matching skills (e.g. Bate, Frowd et al., 2018; Bobak, Bennetts et al., 2016; Bobak, Dowsett et al., 2016; Bobak, Hancock et al., 2016; Davis et al., 2016). These dissociations have generated suggestions that different types of SR might exist, although poor performances on any test may be due to a lack of engagement, distractions or other factors not related to face recognition ability. Indeed, long-term memory for faces in particular may be impacted by a very large number of intervening and uncontrollable variables. As a consequence, some authors have also suggested that to reliably identify SRs, more than one test of face recognition is required (e.g. Noyes et al., 2017). The inclusion of the GFMT also allowed assessment of whether outcomes differed if one (CFMT+), or two (CFMT+ and GFMT) tests were employed to assign participants to SR and control groups.

As with previous research investigating SRs, a disproportionately high number of participants in the current research achieved SR criteria (e.g., Belanova et al., 2018; Satchell, Davis, Julle-Danière, Tupper, & Marshman, 2019). Many more scored slightly below our SR threshold. This is likely due to interest in the topic However, restricted range samples can reduce the strength of correlations (e.g. Goodwin & Leech, 2006). A three-component strategy was therefore employed to reduce the impact of this recruitment bias. First, with the inclusion of all participants, correlations examined relationships between all test outcomes. Second, the performances of SRs and ‘typical-range ability’ controls were compared. Third, individual analyses compared performances of each SR against the mean of controls. This generated an estimate of the proportion of short-term memory test identified SRs who would also be considered SRs at simultaneous face matching and long-term face recognition.

Experiment 1

As police working in SR units may be required to identify suspects over delays of 40 days or more, in Experiment 1, participants completed the CFMT+ (Russell et al., 2009), and the GFMT (Burton et al., 2010) prior to starting a 10-trial *Long-Term Face Memory Test* (LTFMT10) with retention intervals of 1 to ≥ 56 days. As this research was exploratory, exact intervals were 1 day, 7 days, 14 days, 28 days and 56 days, in case task difficulty with longer delays was too great. In Phase 1 of the LTFMT10 (learning phase) participants viewed ten 60s target actor videos. In Phase 2 (test phase) they identified the targets from ten 6-person target-present *hybrid-video* line-ups. With these line-ups, each member's video played sequentially, while still images of the others remained visible. The aim was to reduce floor effects as simultaneously presented line-ups generate higher accuracy than sequential line-ups (e.g. Mickes, Flowe, & Wixted, 2012), while moving images can also enhance unfamiliar face recognition (e.g. O'Toole, Roark, & Abdi, 2002). This may be due to additional cue availability in multiple video frames, rather than movement itself (although see Lander, Christie, & Bruce, 1999).

A sub-set of participants were also allocated to SR and typical-range ability control groups (i.e. within 1 standard deviation of the typical population mean) based on their CFMT+ scores. Previous research has found positive relationships between simultaneous face matching and short-term face memory (e.g. McCaffery et al., 2018), and between short-term and long-term face memory (e.g., Bindemann et al., 2012; Davis & Tamoytė, 2017), with longer retention intervals resulting in lower accuracy (e.g. Deffenbacher et al., 2008). Similar effects were predicted here.

Method

Design

In Experiment 1, participants first completed the CFMT+ (Russell et al., 2009) and the GFMT (Burton et al., 2010), and another short-term face memory test [1], before viewing ten 60s target-actor videos in Phase 1 of the LTFMT10. After random retention intervals of 1, 7, 14, 28, or 56 days, they received invites to Phase 2 of the LTFMT10 during which they viewed ten 6-person target-present hybrid-video line-ups. The dependent variables were correct line-up identification rates (hits), incorrect foil identification rates (foil IDs), and incorrect rejections of the line-ups (misses). A correlational design examined the relationships between performances on each test.

With a sub-set of participants, a 2 (*group*: SRs, controls) x 5 (*actual retention interval*: 1-6, 7-13, 14-27, 28-55, 56+ days) between-subjects design evaluated outcome differences, while individual analyses (Crawford, Garthwaite, & Porter, 2010) compared performances of each SR against the control mean.

Materials

Cambridge Face Memory Test: Extended (CFMT+: Russell et al., 2009).

This 102-trial test is an updated version of the standard 72-trial CFMT (Duchaine & Nakayama, 2006). Participants' short-term learning and recognition of six white-Caucasian male hairstyle-cropped target faces is tested in an increasingly difficult three-alternative forced-choice paradigm across four blocks.

Bobak, Pampoulov et al.'s (2016) participants may be the most representative UK sample to take this test (n = 254, M = 70.72, SD = 12.32). Consistent with the expected recruitment bias, with strong effect sizes, a one-sample t-test revealed that participants in

[1] The data from this test are available in the online repository. They are not reported here.

Experient 1 significantly outperformed this standard (M = 86.62, SD = 9.52), t(596) = 40.87, p < .001, *Cohen's d* = 1.45 (see Figure 1).

Glasgow Face Matching Test (short version) (GFMT) (Burton et al., 2010).

This 40-trial test contains simultaneously presented pairs of white-Caucasian male and female facial images. Twenty trials are matched, in which a correct response is 'same' (recorded as a 'hit'). Twenty are mismatched and require a "different" response ('CR').

Current participants (M = 37.3 (93.3%), SD = 2.5), significantly outperformed Burton et al.'s norms (M = 32.5 (81.3%), SD = 9.7), t(596) = 47.83, p < .001, *Cohen's d* = 1.48 (see Figure 1).

Figure 1 here

Long-Term Face Memory Test (10-trial: LTFMT10).

For Phase 1 of this test, ten videos depicting white-Caucasian target males (n = 5) and females (n = 5) in good lighting were sequentially displayed for 60s each. The actions made by the targets, their clothing and background environment information differed to assist discrimination. For instance, some actors walked towards the camera inside a building, others were depicted outdoors, one sat behind a table, another stood behind a desk. One was at a golfing range. Each video clip displayed approximately 20-30s close-up facial views (see Figure 2).

Figure 2 here

For Phase 2, the 10 actors, wearing different clothing to Phase 1, were filmed by an experienced police officer at a London police station to create PROMAT™ video line-ups

(Promat Envision International, Nelson, Lancashire, UK). A 15-sec head-and-shoulders video was filmed of each actor, and following Police and Criminal Evidence Act (1984) Codes of Practice that govern identification procedures in England and Wales, the officer selected eight foils of the same age, ethnicity, and 'position in life' from a volunteer database of over 23,000 videos (see Davis, Maigut, Jolliffe, Gibson, & Solomon, 2015, for a video depicting the PROMAT system).

Normally, PROMAT line-ups contain nine members, each displayed sequentially for 15s. However, pilot testing (n = 40: none were SRs) in which participants viewed the 10 Phase 1 videos, and after 7-day retention intervals, the 10 9-person line-ups revealed close-to-floor performances. Therefore, to reduce task difficulty, the three foils most commonly selected by pilot participants were excluded from each line-up, being likely the hardest to distinguish from targets. In addition, as simultaneously presented line-ups may generate higher accuracy (e.g. Mickes et al., 2012), a 3 x 2 *hybrid-video* design was created. With these, from top left, to bottom right, each line-up member's video played sequentially (12s each), while a still frontal image of the other members remained visible in an array. The sequence repeated until a response was made. Note: To reduce time demands, 1.5s was cut from the start and end of all videos while all line-up members faced the camera. This change was imperceptible.

Participants

Participants contributed after reading media articles about super-recognition linked to an online anonymised 5-min, - *Could you be a Super-Recogniser Test* [2]. On completion, they were invited to participate in the current research. None were compensated for their time, and

[2] www.superrecognisers.com

all duplicate entries, and participants who had taken any of the tests previously were excluded. In total, 1570 completed the CFMT+, the GFMT, and then viewed Phase 1 of the LTFMT10. However, only 597 (38.0%) completed Phase 2. Reasons for high drop-out are reported below. All participants were included in correlational analyses (n = 597, male = 241, female = 356; aged 16-74 years, M = 35.3, SD = 11.6; White-Caucasian = 505 (64.3%), second largest group (black = 10)).

Inclusion criteria for SR and typical-ability control groups were based on CFMT+ scores. SR threshold was a CFMT+ score (95+ out of 102) in the top 2% (i.e. 2 SD above the mean) of Bobak, Pampoulov et al.'s (2016) representative sample. To reduce the recruitment bias in which a disproportionate number of participants scored fractionally below SR criteria (see Figure 1), typical-ability controls scored within 1 SD of that mean (58-83 out of 102). Participants scoring 84-94 and 57 and below on the CFMT+ were excluded from between-groups analyses. Despite attempts to reduce the recruitment bias, controls still significantly outperformed Bobak, Pampoulov et al.'s (2016) mean on the CFMT+, $t(164) = 4.38$, $p < .001$, Cohen's $d = 0.44$.

Table 1 displays SR and control group demographic information and mean CFMT+ and GFMT performances. Independent-measures t-tests demonstrated significant differences with strong effect sizes, whereby SRs outperformed controls on both tests.

Table 1 here

Procedure

Participants were invited to contribute to research loaded on the Qualtrics platform[3], and asked not to use tablets/mobiles to optimise image size. After providing informed consent, they completed the CFMT+, the GFMT, and the unreported test, and were then provided with debriefing information which included a list of their scores on these tests.

All participants were then immediately invited to participate in the current research project which included the LTFMT10. After providing informed consent for a second time, which included access to scores on the CFMT+ and GFMT and information that they would need to provide e-mail addresses removing anonymity, they were familiarised with requirements by viewing a practice trial. This consisted of a cartoon character Phase 1-type video and immediately afterwards, a cartoon Phase 2-type target-present line-up, with instructions, and options to select a line-up member (or not) and to provide confidence as described below.

Next, participants sequentially viewed the Phase 1 target actor videos. Unless problems were reported, videos could not be replayed. After each video, questions asked whether, a. participants had seen that actor before (none had) and b. whether the video had played properly. If participants clicked 'no' to b, that video was repeated after the final video of 10. There were 372 reported problems out of 15,700 plays (2.36%) (n = 203 participants). Repeat plays had no impact on whether participants completed Phase 2 or not; or on any results reported below ($p > .2$).

After viewing all Phase 1 videos, participants were asked to provide an e-mail address and were warned to watch out for their Phase 2 invite e-mail. As retention interval was random, no information was provided as to when the e-mail would be sent. In total, 1,570 participants completing LTFMT10 Phase 1 were e-mailed Phase 2 invites, with 597 finishing the LTFMT10 (38.0%). Some drop out can be explained from high numbers of rejected e-

[3] www.qualtrics.com

mails – suggesting errors in entering details. A point-biserial correlation examining the relationship between CFMT+ scores and finishing or not (1 = finisher, 0 = non-finisher) was positive and significant, $r(1570) = .16$, $p < .001$. More SRs (48.0%) completed Phase 2 than controls (28.7%), $\chi^2(1, 534) = 19.28$, $p < .001$, Cramer's $V = .190$.

On clicking on the Qualtrics link for Phase 2, the final line-ups were sequentially displayed in the same order as first phase videos (i.e. Actor A's Phase 1 video and Phase 2 line-up was displayed first). Target-actor and foil positions within line-ups were randomised and counterbalanced. In the pilot research described above, participants were provided with the usual police warning that the "culprit may or may not be present in the line-up". This resulted in high incorrect foil ID *and* miss rates. To encourage selections in the final research, participants were correctly informed that "the target-actor is definitely present in each line-up", although an option to reject the line-up was retained. Participants identified targets by selecting a line-up number (1-6), or if they did not recognise the target, they rejected the line-up ('none of the above'). Correct line-up identification rates (hits), incorrect foil identification rates (foil IDs), and incorrect line-up rejections (misses) were calculated. Confidence ratings in identifications were collected immediately after each line-up decision (0%: guessing to 100%: absolutely certain). However, as all participants were told all trials were target-present, which would not match normal eyewitness procedures, as it potentially encouraged guessing, confidence data were analysed in Experiment 2 only [4].

Regardless of retention interval, mean hit rates to each actor on the LTFMT10 were higher than chance and ranged from 68% to 26% (chance levels = 1/6 = 16.7%).

Results

[4] Note: Confidence data are available in the supplementary data.

LTFMT10 retention interval (Median = 14 days, *M* = 24.1, *SD* = 23.9) varied from 1-182 days (see Figure 3); and was skewed (Shapiro-Wilk (597) = 0.83, $p < .001$), so that where appropriate, non-parametric tests were employed. Mean retention interval did not differ between SRs and controls, $t(284) < 1$. Seven participants (SRs *n* = 3; Controls *n* = 0; Excluded *n* = 4) achieved LTFMT10 scores of 10 out of 10 (*n* = 21 scored 0) (Median = 4, *M* = 4.08, *SD* = 0.24). Despite being told that targets were definitely present in the line-up, 15.6% of participant selections were misses (participants responded 'none of the above').

Figure 3 here

Correlational analyses: With the inclusion of all participants, Table 2 presents Spearman's correlation coefficients across all test outcomes. Retention interval did not correlate with GFMT and CFMT+ scores suggesting successful randomisation. While scores on all three face processing tests were as expected significantly positively related, the correlation between scores on the GFMT and the CFMT+ was stronger than the positive correlation between these tests and LTFMT10 hits, and the negative correlation between these tests and LTFMT10 Foil IDs. As predicted, LTFMT10 hit rates were also negatively correlated with retention interval, Foil IDs, and misses.

Table 2 here

Between-groups analyses: To compare performances of SRs and controls, three 2 (group: SR, control) x 5 (retention interval: 1-6 days (actual mean retention interval: *M* = 1.46), 7-13 days (*M* = 7.91), 14-27 days (*M* = 15.37), 28-55 days (*M* = 31.77), 56+ days (*M* = 64.12)) ANOVAs were performed on hits, foil IDs and misses on the LTFMT10. Outcomes

are reported in Figure 3 along with the results of Tukey's post-hoc tests [5]. These analyses provide an estimate of forgetting rates, assuming that the mean hit rates in the 1-6 days retention interval condition represent faces reliably committed to memory.

Significant group main effects were found on LTFMT10 hits, $F(1, 276) = 29.25$, $p < .001$, $\eta^2 = .096$, and foil IDs, $F(1, 276) = 37.12$, $p < .001$, $\eta^2 = .119$, but not misses $F(1, 276) < 1$. SRs made more hits and fewer foil IDs than controls.

Significant retention interval effects were found on hits, $F(4, 276) = 7.61$, $p < .001$, $\eta^2 = .099$, and foil IDs, $F(4, 276) = 6.36$, $p < .001$, $\eta^2 = .084$, but not misses, $F(4, 276) < 1$. Longer retention intervals were associated with reduced hits and increased foil IDs. However, although mean hits in the 1-6 days interval were significantly higher than all other retention intervals, and mean false alarms significantly lower than the 14-27, 28-55, and 56+ day intervals, only a few of the remaining paired comparisons examining the differences between all retention intervals were significant (see Figure 4).

Figure 4 here

There were no significant interactions, $F(4, 276) \leq 1.16$, $p \geq .330$, $\eta^2 \leq .016$, suggesting that the forgetting curves for SRs and controls are of a similar shape, although see Discussion.

Individual level analyses: Modified t-tests for single cases (Crawford et al., 2010), individually compared the scores of each SR against the control mean on the CFMT+ (out of

[5] Some authors argue that to rigorously allocate SRs to groups, two or more tests are required. Following Satchell et al. (2019), these three ANOVAs were therefore repeated with the exclusion of SRs ($n = 41$) achieving less than maximum on the GFMT (GFMT = 40), and controls ($n = 89$) scoring more than 1 SD outside the typical population mean on the GFMT (GFMT = 28-36 out of 40). These ANOVAs generated virtually identical effects as the main analyses reported above, albeit with larger between-group effect sizes for hits, $F(1, 120) = 18.11$, $p < .001$, $\eta^2 = .131$, foil IDs, $F(1, 120) = 41.38$, $p < .001$, $\eta^2 = .256$, and misses, $F(1, 120) = 7.29$, $p = .008$, $\eta^2 = .057$. Indeed, unlike in the main text, miss rates became significant using these criteria. SRs ($M = 0.19$; $SD = 0.22$) made more misses than controls ($M = 0.12$; $SD = 0.16$).

102), the GFMT (out of 40), and LTFMT10 (hits) within each retention interval (see Figure 5).

Figure 5 here

Not surprisingly given inclusion criteria, all SRs (n = 165, 100%) scored significantly higher than the control mean on the CFMT+, $t(165) = 3.02\text{-}4.08$, $p < .05$, one-tailed, $z = 3.03$ (95% CI: 2.67-3.39) – 4.09 (95% CI: 3.63-4.56).

Most SRs exceeded the control GFMT mean (n = 111, 91.7%). However, even maximum scores of 40 out of 40 were not significantly different from the control mean. Noteworthily, ten SRs (8.3%) scored below the control GFMT mean, though non-significantly.

For the LTFMT10, SRs were compared to controls from the matched retention interval group described in Figure 2. The scores of 95 SRs (78.5%) exceeded the associated retention interval control mean, although only 31 (25.6%) significantly outperformed controls, $t(121) = 1.78\text{-}2.74$, $p < .05$, one-tailed, $z = 1.81$ (95% CI: 1.23-2.38) -2.77 (95% CI: 2.11-3.43). A further 26 SRs scored below control means, two – significantly ($p < .05$). Importantly, the worst SR performers on the LTFMT10 were not necessarily those who scored poorly on the GFMT.

Discussion

In Experiment 1, as expected, correlational analyses with the inclusion of all participants demonstrated that longer retention intervals were associated with significantly fewer LTFMT10 hits, and more foil IDs. With stronger effect sizes, as a group, SRs were significantly more accurate than controls regardless of retention interval condition.

Importantly, there was no interaction between group and retention interval. As such, the shape of SRs and controls' forgetting curves appears roughly similar, and SRs' superiority over controls after 56 days is of approximately the same degree as after 1 day. It is important to note, however, that genuine interaction effects typically require larger sample sizes to provide adequate power to detect them (Maxwell, Delaney, & Kelley, 2018), and therefore the presence of genuine interaction effects cannot be dismissed. Numbers of SRs in some retention interval conditions were low (< 20), resulting in low power (.23) to find small effects in Experimen1. Experiment 2 recruited a far larger participant sample to address this limitation.

Despite the low numbers of SRs in some conditions, between-group outcomes were matched by the individual analyses. All SRs (n = 121, 100%) significantly exceeded the control group mean on the CFMT+ and most SRs exceeded the GFMT (91.7%); and LTFMT10 (78.5%) control means. However, only a minority significantly outperformed the control group on the LTFMT10 (25.6%). This may be due to low discriminatory power from only including 10 trials in the LTFMT10, as well as the bias to recruit better-than-typical face recognisers – the control group here significantly outperformed past research norms on the CFMT+ and GFMT. This may partly be because a higher proportion of controls (71.3%) than SRs (52.0%) dropped out after completing Phase 1. It is possible many of these controls were aware that their LTFMT10 scores would be poor and declined to continue. All participants received feedback in the form of their scores after the CFMT+ and GFMT, from which controls would have been able to infer they were not exceptionally good at this task – a possible demotivating effect. On the other hand, the investment in time of those who did finish was high, and it would be surprising for someone to take part but not try. Nevertheless, a few SRs also performed below the control mean on the GFMT (8.3%), supporting proposals

for dissociated perceptual and memory-based processing components at the highest levels of ability (e.g., Bate, Frowd et al., 2018).

In summary, Experiment 1 demonstrated that superiority on the short-term CFMT+ is associated with superiority at the long-term recognition of faces, while the forgetting curve of SRs and controls was roughly similar. However, participants were correctly informed Phase 2 trials were all target-present, meaning participant deductions as to the most likely target in the line-up may have contributed to outcomes. Furthermore, most forgetting happens in the first 24 hours (Deffenbacher et al., 2008), and no performance baseline with virtually no LTFMT retention interval was included to establish a baseline of performance and to measure initial decline. Finally, a large number of participants (62%) failed to finish the LTFMT10. Experiment 2 addressed these factors.

Experiment 2

In Experiment 2, participants completed an extended 20-trial (LTFMT20) version of Experiment 1's 10-trial LTFMT10. To maintain Experiment 1's 10-min Phase 1 learning time, and to avoid elevated participant drop out risk, learning time for each Phase 1 video in Experiment 2 was reduced by half, to 30s. After random retention intervals (immediate, 1 day, 7 days, 14 days, 28 days), all participants viewed ten target-present and ten target-absent 3 x 2 counterbalanced simultaneous photo line-ups. The use of photos instead of videos was expected to result in lower rates of accuracy than Experiment 1. However, it also reduced Phase 2 time demands. From an applied perspective, police are also more likely to first view suspects in CCTV stills in briefings, before deciding whether to access the original crime scene videos. This would be more likely if they believed they recognised the suspect from the still image. As such, it is important to measure factors that might impact on performances in

real world situations. Finally, all participants in Experiment 2 had previously taken the GFMT (Burton et al., 2010) and the CFMT+ (Russell et al., 2009) in unpublished research, and had requested invites to future research. Recruiting motivated volunteers was also expected to reduce dropout.

In Experiment 2, identification decision confidence was also analysed. Signal detection theories of decision making and confidence processing (e.g., Green & Swets, 1966; Macmillan & Creelman, 1991) suggest response and response confidence in recognition memory tasks derive from shared evidential bases. As such, conditions amenable to higher identification decision accuracy (i.e. longer exposure and good views of targets in learning phases; shorter retention intervals), should also generate higher confidence. The opposite should apply if identification conditions are poor. In support, most research finds identification accuracy and confidence is calibrated (e.g., Mickes et al., 2012; Sauer et al., 2010; Sporer, Penrod, Read, & Cutler, 1995, Wixted & Mickes, 2018; for a review see Sauer & Brewer, 2015). Indeed, Sauer et al. (2010) found that regardless of retention interval, higher confidence was associated with stronger diagnosticity. However, compared to when identification decisions were made immediately, there was a tendency for participants to be over-confident in their identification decisions after 20-50 days, reducing diagnosticity. Superior face recognition ability is also associated with higher confidence in accurate identifications, and lower confidence in misidentifications (e.g., Grabman, Dobolyi, Berlovich, & Dodson., 2019; see also Davis et al., 2018), and the current design allowed for measurement as to whether this relationship persists over longer retention intervals.

Diagnosticity may be confounded with response bias, or participants' tendency to choose (or not) from a line-up (e.g., Wixted & Mickes, 2012). Therefore, for the current research, confidence-based accuracy in SRs and controls was plotted using Receiver Operating Characteristic (ROC) curves. ROCs plot the probability of a correct identification

of a target in target-present line-ups (y axis) against the probability of an incorrect foil identification (false alarm: FA) in target-absent line-ups (x axis) at each confidence level. The area under the ROC (AUC) is independent of response bias (Green & Swets, 1966). Here, a partial area under the curve (pAUC) was reported because the false alarm rate was limited by the number of lineup members (n = 6), meaning the curve cannot extend across the entire x-axis (e.g., see also Mickes et al., 2012).

In eyewitness research in which the ground truth of 'guilt' is known, if correct identifications of guilty suspects (true positives) are plotted on the y-axis and incorrect identifications of foils (false alarms) on the x-axis, then the AUC reflects the ability of witnesses to distinguish between guilty and innocent suspects, independently of willingness to identify anyone. In the current research, it was predicted that SRs would display larger AUC than controls at each retention interval, reflecting better discrimination of targets from foils.

Consistent with previous findings, a positive relationship was also expected between CFMT+, GFMT and long-term face recognition test performances. Retention interval was predicted to have a negative impact on LTFMT20 accuracy, with the largest retention interval effect sizes expected for target-present than target-absent trials, indicative of increased forgetting of targets. Regardless of retention interval, SRs were predicted to generate more target-present hits and target-absent correct rejections (CRs) than controls.

Method

Design

In the LTFMT20 Phase 1, participants viewed twenty 30s target-actor videos. In Phase 2 (immediately, 1 day, 1 week, 2 weeks, or 28 days later), they attempted to identify the actors from 10 target-present and 10 target-absent six-person photo line-ups. Target-presence order was counterbalanced. After making an identification decision they provided confidence ratings in that decision (0%: guessing to 100%: absolutely certain). A correlational design examined relationships between scores on the CFMT+, the GFMT, as well as LFMT20 retention interval, hits and CRs. A mixed 2 (*group*) x 5 (*retention interval*: < 1 day, 1-6 days, 7-13 days, 14-28 days, >28 days) x 2 (*target-presence*: target-present *vs.* target-absent) design, with target-presence as the within-subjects factor, also evaluated each outcome. At each retention interval, partial ROC analyses additionally compared SR's and control's probability of a correct identification of a target in target-present line-ups against the probability of a foil identification in target-absent line-ups at each confidence level using a stratified bootstrap method. Finally, individual analyses again compared test outcomes of each SR against the mean performances of controls.

Materials

Long-Term Unfamiliar Face Memory Test 20 (LTFMT20)

Procedures for this test mostly matched those described in Experiment 1, except trial numbers were doubled, and half the trials were target-absent. In Phase 1, participants viewed 20 30s target actor videos, ten of which had been used in Experiment 1's LTFMT10. These clips, originally 60s, were cut at appropriate places to ensure that close-up views of each target were retained without impacting action continuity. Videos of ten new targets were additionally selected for Experiment 2 from an image database of about 500 students. Experiment 1's LTFMT10 targets comprised the odd-numbered trials in the LTFMT20 (e.g.

1, 3, 5 etc.). The ten new Phase 1 videos and Phase 2 line-ups comprised the even-numbered trials. The new Phase 1 videos depicted targets from close-up, moving their heads from side-to-side in a similar manner to the PROMAT line-ups described in Experiment 1.

In Phase 2, participants were presented with 20 3 x 2 photo line-ups, each consisting of two stills of each line-up member (frontal view, right profile). For the 10 new Phase 2 line-ups, the targets wore different clothing from Phase 1. All new line-up photos were taken against the same blue background. To ensure a balanced design with each target associated with a target-present and target-absent line-up, Phase 2 had two randomly allocated versions.

Overall, regardless of retention interval, individual trial accuracy ranged from 82.0% to 13.0% for hits in target-present trials, and 51.0% to 27.0% for CRs in target-absent trials.

Participants

New volunteer participants who had completed the CFMT+ (Russell et al., 2009) and the GFMT (Burton et al., 2010) once only previously were invited to Experiment 2. Overall, 1,683 participants completed Phase 1 of the LTFMT20. Drop out was far less than Experiment 1 as 270 failed to finish the LTFMT20 (16.0% dropped out). Unlike in Experiment 1, Experiment 2's participants were recruited via email invitations and had fewer tests to complete, which potentially explains why a smaller proportion of participants failed to finish. Substantially more females than males completed the LTFMT20 (n = 1,421; male = 445, female = 972, other = 3; white-Caucasian = 1236 (86.9.%), aged 16-86 years, M = 39.87, SD = 12.14)[6].

Figure 6 here

[6] Note: demographic information was missing for some participants (gender n = 1; age n = 3; ethnicity n = 3)

SR and control groups were based on similar criteria as Experiment 1. Table 3 depicts the final group demographic information as well as mean CFMT+ and GFMT scores with independent t-tests comparing groups. As with Experiment 1, despite attempts to reduce the impact of the recruitment bias, controls still significantly outperformed Bobak, Pampoulov et al.'s (2016) sample on the CFMT+, $t(414) = 14.31$, $p < .001$, Cohen's $d = .46$ (Figure 6).

Table 3 here

Procedure

Invitees were e-mailed a Qualtrics (www.qualtrics.com) link and asked not to use tablets/mobiles to optimise image size. After providing informed consent, participants viewed Experiment 1's LTFMT cartoon practice trial, followed by the twenty target-actor videos (30s each) in Phase 1. As with Experiment 1, if participants reported problems with video play by clicking on an icon, they could view that video again at the end of Phase 1. There were 628 replays out of 28,460 videos viewed (2.21%), which had no effect on any reported results ($p > .2$). Participants automatically received Phase 2 URL link e-mails at random retention intervals (immediately, 1 day, 7 days, 28 days), although not all took part immediately on receiving the e-mail. In Phase 2, they were asked to identify recognised actors by selecting a line-up number (1-6), or to reject the line-up ('none of the above'). They were not provided with any other instructions. Finally, they were debriefed.

Unlike Experiment 1, a point-biserial correlation examining the relationship between CFMT+ scores and finishing or not (1 = finisher, 0 = non-finisher) was not significant, $r(1683) = -.04$, $p = .142$. Nevertheless, more SRs (98.1%) completed Phase 2 than controls (83.8%), $\chi^2(1, 563) = 27.91$, $p < .001$, Cramer's $V = .223$.

Results

LTFMT20 retention intervals varied from almost immediately to 50 days (Median = 7.6, *M* = 13.3, *SD* = 12.2) (see Figure 7); and was skewed (Shapiro-Wilk (1421) = 0.842, $p < .001$). Overall, mean retention intervals were significantly *longer* for SRs than controls, $t(714) = 2.04$, $p = .041$, Cohen's $d = .15$. The highest total score by any participant ($n = 1$) was 19 out of 20 ($n = 3$ scored 0) (Median = 8.0, *M* = 8.2, *SD* = 3.3) (Figure 7).

Figure 7 here

Correlational analyses: Table 4 presents correlation coefficients between test outcomes and retention interval in the LTFMT20. As in Experiment 1, there was no evidence of retention interval differing by ability as correlations with CFMT+ and GFMT scores were close to zero. Accuracy rates on all face processing tests (CFMT+, GFMT, and LTFMT20 hits, and CRs) were positively and significantly related; although the coefficients between the CFMT+, and the GFMT, were stronger than those between these tests and the LTFMT20 outcomes. Furthermore, as hypothesised, in comparison to those between the CFMT+, GFMT, and LTFMT20 hits, weaker correlations were revealed between the same two tests and CRs on the LTFMT20. Finally, as expected, retention interval negatively correlated with LTFMT20 hits indicative of forgetting of targets, but not LTFMT20 CRs.

Table 4 here

Between-group analyses: To compare performances of SRs and controls, two separate 2 (group: SRs, controls) x 5 (retention interval condition: < 1 day (actual mean retention

interval: M = 0.33), 1-6 days (M = 5.44), 7-13 days (M = 8.80), 14-27 days (M = 24.99), 28+ days (M = 30.86)) ANOVAs were conducted on LTFMT20 hits and CRs (see Figure 8).

Unlike Experiment 1, no 14-day Phase 2 e-mail invites were sent. However, large numbers of participants completed Phase 1 between 14-27 days, allowing this retention interval category to be included in order to match Experiment 1's results. All analyses were first conducted with the addition of the two Phase 2 counterbalanced conditions as a variable. Hit rates (M = 0.49, SD = 0.21 *vs.* M = 0.39, SD = 0.22) were significantly higher in one counterbalanced condition than the other ($p < .001$). However, there were no effects or interactions involving other variables ($p > .2$), and these data were collapsed.

LTFMT20 hits: Significant main effects of group, $F(1, 706) = 86.39$, $p < .001$, $\eta^2 = .109$, and retention interval, $F(4, 706) = 41.83$, $p < .001$, $\eta^2 = .192$, showed that as predicted, SRs (M = 0.51, SD = 0.23) outperformed controls (M = 0.38, SD = 0.20), and that hit rates were higher with shorter retention intervals. Post hoc comparisons showed hits were successively reduced at each retention interval, although not all paired comparisons were significant (see Figure 8). Consistent with Experiment 1, there was no significant interaction, $F(4, 706) < 1$, $p = .830$, $\eta^2 = .003$.

Figure 8 here

LTFMT20 CRs: A significant main effect of group, $F(1, 706) = 23.65$, $p < .001$, $\eta^2 = .032$, showed that SRs (M = 0.43, SD = 0.27) outperformed controls (M = 0.33, SD = 0.25). Neither the effect of retention interval, $F(4, 706) < 1$, $p = .699$, $\eta^2 = .003$, nor the interaction was significant, $F(4, 706) < 1$, $p = .964$, $\eta^2 = .001$.[7]

[7] Between-group analyses with the additional GFMT inclusion criteria described in Experiment 1 and Satchell et al. (2019) generated similar results as those reported here, albeit with slightly different effect sizes. The main effects of group (SR n = 135; controls n = 239) generated stronger effect sizes for both hits and CRs, while smaller effect sizes were found for the main effect of retention interval in hits.

ROC analysis: Figure 9 shows ROC curves for SRs (uninterrupted lines) and controls (dashed lines) across five retention intervals with confidence-based hits plotted as a function of confidence-based FAs (calculated from 1-CRs) using the method described by Mickes et al. (2012; see also Wixted & Mickes, 2018). The 10 responses on target-present trials and 10 responses on the target-absent trials by each participant were pooled for these analyses.

Figure 9 here

The curves produced for SRs and controls display five markers/pointers, one for each confidence range, so that the first marker in the left top corner displays diagnosticity ratio derived from hits and FAs for 90-100% confidence range. The second marker displays diagnosticity ratio for 70-100% confidence range and so on. ROC analyses calculated the partial AUC for SRs and controls under each retention interval. Visual inspection shows that at all retention intervals, SR's identification performances were closer to the top left corner indicative of stronger precision. Precision is also reduced with longer retention intervals for both groups. SRs AUC after 28 days is roughly similar to that of controls in the immediate condition (D = .986; p = .324). The AUC, indicative of better identification discrimination, was significantly greater in SRs than controls for all retention intervals (p < .01).

Individual analyses: As with Experiment 1, modified t-tests for single cases (Crawford et al., 2010), individually compared SR's scores against the control mean on the CFMT+, GFMT, as well as LTFMT20 hits, and CRs within each matched retention interval (see Figure 10 depicting 95% CIs of the estimated proportion of the general population each SR would be expected to exceed on each measure). As there were significant differences between the two counterbalanced LTFMT20 conditions, SRs were compared against the controls in their condition only.

Not surprisingly given inclusion criteria, all SRs (n = 301, 100%) scored significantly higher than the control mean on the CFMT+, $t(415) = 3.09\text{-}4.18$ $p < .05$, one-tailed, $z = 3.09$ (95% CI: 2.86-3.32) – 4.19 (95% CI: 3.89-4.49). Most SRs (n = 280, 93.02%) outperformed controls on the GFMT, 135 (44.9%) significantly, $t(415) = 1.59$ $p < .05$, one-tailed, $z = 1.59$ (95% CI: 1.44-1.73) (control group GFMT mean and variance (*SD*) was slightly smaller in Experiment 2 than Experiment 1 allowing a maximum score to be significant in this experiment only). One SR scored significantly below the control mean ($p < .05$).

Figure 10 here

For the LTFMT20 hits, 225 SRs (74.8%) exceeded the associated retention interval control mean, with 84 (27.9%) significantly outperforming controls, $t(415) = 1.54\text{-}4.27$, $p < .05$, one-tailed, $z = 1.55$ (95% CI: 1.17-1.93) – 4.29 (95% CI: 3.61-4.97). A further 76 SRs (25.2%) scored below the control mean, eight significantly ($p < .05$).

For the LTFMT20 CRs, 176 SRs (58.5.%) exceeded the associated retention interval control mean, although only 54 (17.9%) significantly outperformed controls, $t(415) = 1.51\text{-}2.78$, $p < .05$, one-tailed, $z = 1.52$ (95% CI: 1.12-1.91) -2.79 (95% CI: 2.32-3.26). A further 125 SRs (41.5%) scored below control means, although no differences were significant.

Discussion

The target-present results of Experiment 2 closely matched those of Experiment 1 and hypotheses. For all participants, the highest hit rates were associated with the shortest retention interval (< 1 day). Hit rates at this interval were significantly higher than the other retention intervals. There were also significant, positive, but moderate correlations between

scores on the face matching (GFMT) and short-term face recognition tests (CFMT+), and weak to moderate correlations between these and LTFMT20 hits. Regardless of retention interval, SRs also made significantly more LTFMT20 hits than controls.

The results for hits in target-present trials contrast with those for the target-absent trials, as there was virtually no correlation (r = -.04) between retention interval and CR rates. Significant positive correlations between GFMT and CFMT+ scores and LTFMT20 CRs were also far weaker than those in target-present trials. However, SRs still made significantly more CRs than controls, suggesting that in comparison to controls, SRs are better able to rule out that faces have not been seen before. SR's superiority was also reflected in the ROC analyses which considered participants' confidence-based performance. SRs displayed a greater AUC compared to controls in all retention interval conditions, suggesting that super-recognition is more diagnostic of accurate line-up identification and confidence.

Note that even with far greater power (.77) to detect small effect sizes (.20) than in Experiment 1, and regardless of participant inclusion criteria (i.e., CFMT+ and GFMT based classification, or CFMT+ alone), there was no interaction between group and retention interval. Interaction effect sizes were very small. Therefore, similar to Experiment 1, the shape of SRs and controls' forgetting curve for faces appears roughly similar. This is further elaborated on in the General Discussion.

General Discussion

The two experiments described here demonstrate that most SRs who display exceptionally good short-term face memory possess better-than-typical long-term face memory as well. Retention intervals varied from virtually none in Experiment 2, to over 56 days in Experiment 1. With small effect sizes, SRs LTFMT hit rates in target-present trials

were significantly higher than typical-ability controls. Combining results from both experiments, individual analyses show that the hit rates of most SR's exceeded the control mean of the LTFMT in their associated retention interval, although only just over one-quarter of comparisons were significant. With smaller effect sizes, increasing retention intervals negatively impacted performance, albeit not all comparisons were significant, with effects being roughly similar for SRs and controls.

In Experiment 2's target-absent trials, SRs also made more CRs than controls, although while most SRs individually exceeded the control mean on the individual analyses, only a few comparisons were significant. Importantly, in both experiments, there was no significant interaction between group and retention interval. This suggest SRs' long-term superiority is likely to be a direct result of their short-term memory superiority (whether it is due to enhanced encoding, enhanced retrieval, or both), which is sustained over longer intervals. The lack of an interaction suggests that SRs and controls have similar forgetting curves for faces in memory. It is possible that SRs possess more effective long-term retention which was not detected in this study. For instance, we did not control for the potentially numerous interfering factors SRs and controls will have encountered between Phase 1 encoding and Phase 2 recognition. Ideally, future research should attempt to control for individual levels of interference, although over longer intervals, control over conditions will inevitably become far harder.

It is perhaps not surprising that only a few SRs *significantly* made more hits than controls on the LTFMT. With 10 target-present trials, the LTFMT in both experiments will have had low discriminatory power to distinguish between performers of different standards. Furthermore, both experiments suffered from a recruitment bias to attract participants with better face recognition abilities than would be expected from a more representative sample. This reduction of range will have likely reduced between-group effect sizes and correlations

between variables. Steps were taken to reduce the impact of this bias by creating typical-range ability control groups based on scores within 1 SD of the mean on the CFMT+ (Russell et al., 2009) of a highly population-representative UK sample (Bobak, Pampoulov et al., 2016). Despite this, with medium effect sizes, control groups in both experiments still significantly exceeded Bobak et al.'s sample mean.

Between-test performance inconsistencies

Despite moderate to strong positive correlations in both Experiment 1 (r = 0.42) and 2 (r = 0.47) between CFMT+ and GFMT scores of a roughly similar strength to those in previous research (e.g. r = 0.45, McCaffery et al., 2018), a few SRs (1 significantly) also achieved scores below the control mean on the GFMT, suggesting the impact of dissociations between perceptual and memory-based mechanisms (see also Bobak, Bennetts et al., 2016; Bobak, Dowsett et al., 2016; Bobak, Hancock et al., 2016; Davis et al., 2016). Accurate face matching mainly relies on a facial feature-by-feature comparison strategy (e.g. Megreya & Burton, 2006); whereas face memory draws more on holistic or whole face mechanisms (e.g. Tanaka & Farah, 1993; Tanaka, Heptonstall, & Campbell, 2019). A proclivity to employ a holistic face processing style has been associated with superior face memory (e.g., DeGutis et al., 2013; Wang et al., 2012; although see Konar, Bennett, & Sekuler, 2010). These results may simply imply that some SRs may use an inefficient holistically based strategy when completing face matching tasks. Further research is required to investigate this proposal.

The failure of some SRs to outperform controls on the LTFMT may also reflect a different approach to encoding faces during the learning stage. Indeed, face recognition performance is associated with enhanced or deeper encoding (e.g., Marzi & Viggiano, 2010), and the reliance on holistic processing during face encoding similarly contributes to long-

term recognition (Tanaka et al., 2019). While previous SR research provides evidence of some SR's enhanced pictorial encoding (Belanova et al., 2018) and enhanced holistic processing (Bobak, Bennetts et al., 2016), not all SRs display the same pattern of results. It is therefore possible that this heterogenous pattern of encoding and holistic processing during the learning stages contributes to the heterogenous pattern of long-term face recognition performance observed here.

The failure of some SRs to achieve a score above the mean of controls on the LTFMT may also reflect a dissociation between short- and long-term elements of superior face memory. Indeed, some SRs ($n \approx 5$) contacted the researchers stating that although they found the short-term memory tests easy, they struggled with longer term face recognition.

It is also noteworthy that some between-test incontinency is to be expected, and may also reflect varying levels of engagement, distractions, fatigue, or a range of factors unrelated to face recognition ability. In online studies such as the current research, control over conditions is far weaker than in a lab, and therefore a range of extraneous factors may have impacted results. On the other hand, recent research demonstrates that experimenter presence in a laboratory can reduce visual memory retrieval performances, suggesting that sometimes, internet-based studies may better capture abilities (e.g. Souza, Rerko, & Oberauer, 2015). Nevertheless, the results of the current research suggest that recruitment criteria for SR roles in policing and security, should best be based on achieving test scores measuring a range of different skills (see Noyes et al., 2017), rather than high performances on a single short-term face memory test such as the CFMT+ (Russell et al., 2009).

Between-test performance consistency

Regardless of potential dissociations between different elements of superior face recognition ability, a few participants did achieve highly superior performances on all three tests (see Figures 5 and 10). Nevertheless, even in Experiment 2's immediate condition, SR's accuracy on the LTFMT20 was well below ceiling. Mean SR hit rates in this shortest retention interval were 67%, CR rates were 44%. This might imply poor SR reliability. However, task demands were very high, being the equivalent of 20 typical eyewitness identification experiments in one session. Additional factors may have impacted outcomes.

First, the LTFMT Phase 1 videos displayed actors from different viewpoints providing varying levels of full body, gait and facial movement information. The Phase 2 line-ups displayed head-and-shoulders information only. As no information was provided in advance as to Phase 2 task demands, it is possible that some participants in Phase 1 concentrated on the actions in the videos or tried to learn non-face features (e.g. full bodies/clothing). Such a strategy would have reduced performances.

Second, in Experiment 2, all line-ups were photos, which will likely have increased task difficulty as no movement matching that of the Phase 1 videos could be extracted.

Third, police worldwide usually warn eyewitnesses that "line-ups may or may not contain the target". No such warning was given in Experiment 2, although in Experiment 1 participants were correctly told all line-ups were target-present. The unbiased police warning of target presence reduces misidentifications of innocent suspects and foils (e.g. Clark, 2012). However, it often also induces a response bias to be cautious, reducing hit rates as well. It is possible that if such a warning had been provided to participants, outcomes may have differed. Given the normal impact of the warning, CR rates would have been most likely increased, albeit at the possible cost of reduced hit rates.

There were, of course, differences between SRs and controls that may provide an alternative explanation for group effects. Specifically, SRs were more likely to finish both

experiments, achieve higher scores on the CFMT+ and the GFMT, and to assign higher confidence values on the LTFMT. As such, it could be argued that rather than possessing enhanced face recognition ability, SRs are simply more competitive or motivated, and therefore try harder at tests of this type. This proposal would, however, ignore the growing body of literature demonstrating that SRs do not always outperform controls at tasks that do not involve the recognition of faces (e.g. flowers recognition: Davis et al., 2016; house recognition: Bobak, Bennetts et al., 2016; personality in faces: Satchell et al., 2019). It would seem strange that groups of different SRs recruited to different research projects would only be motivated to try hard on face tests, and not on other tests presented in the same research.

Limitations

There are some limitations of this research that should be acknowledged. All tests employed white-Caucasian adults as stimuli. Future research could examine effects with actors of different ethnicities and ages, to measure the impact of the cross-age (e.g. Wiese et al. 2013) and cross-ethnicity effects (Meissner & Brigham, 2001). Robertson, Black, Chamberlain, Megreya, and Davis (2019), and Belanova et al. (2018) found that SRs outperformed controls at simultaneous other-ethnicity face matching, and other-age recognition respectively, and it might be predicted that these advantages would transfer to longer-term recognition as well.

In addition, from a theoretical perspective, only long-term memory of the human face was tested here. Previous research (e.g. Bobak, Bennetts et al., 2016) demonstrates that superior face recognition ability is not always dissociated from object recognition ability, indicating possible domain-general mnemonic enhancements. Future research should therefore compare long-term memory for faces with tests assessing other classes of visual

stimuli to better control for the confounding effects of generalized superior memory processes.

Body information may also enhance identification over faces displayed alone (e.g. Burton, Wilson, Cowan, & Bruce, 1999; Noyes, Hill, & O'Toole, 2017; Robbins & Coltheart, 2015), although body and face matching performances may not necessary correlate (Noyes et al., 2017). For policing and security, it might be advantageous to deploy staff with exceptional face *and* body/gait recognition skills, particularly for roles reviewing CCTV footage, or surveillance and viewing suspects in real life. Recognition of persons of interest for instance in CCTV footage would not necessarily be limited to faces alone and as such, future research could investigate this with whole body line-ups perhaps using Experiment 1's hybrid-video style to capture idiosyncratic movements and gait.

The CFMT+ used here as the primary assessment of SR ability has been commonly employed in this type of research, although utility has been criticised (e.g., Bate, Bennetts et al., 2018; Bate, Frowd et al., 2018; Esins, Schultz, Stemper, Kennerknecht, & Bulthoff, 2016). For instance, the CFMT+ employs only target-present trials, limiting the investigation of participants' face processing skills (i.e. it is not possible to generate data of correct rejections), while the highly controlled greyscale cropped images depicting no hairstyle do not reflect realistic daily face recognition scenarios (Bate, Bennetts et al., 2018). The current research demonstrates that CFMT+ scores do predict longer term face recognition performance, although alternative face memory test designs may be more suitable. On the other hand, between-group outcomes were virtually identical when a second test (the GFMT) was used to assign SRs and controls to groups in both experiments. Although this caused a reduction in participant numbers in both groups, effect sizes were increased, suggesting that a multiple test strategy may be best practice for the selection of SRs to critical roles in organisations such as policing.

Finally, it could also be argued that the modest differences in performances between SRs and controls on the LTFMT in both experiments, despite being significant, might not generate large gains if transferred to applied settings such as policing. However, if a single SR's superior ability to identify suspects over longer periods of time provides the first link in an investigation chain for a highly serious crime (i.e. murder), that might otherwise remain undetected, the positive impact on society would be hard to quantify.

Conclusions

The results from the two experiments reported here were the first to demonstrate that a majority of participants with outstanding short-term face memory ability – the so-called SRs, can sustain these skills over longer-term retention intervals. However, there were exceptions, and some SRs who generated exceptional short-term face memory and simultaneous face matching scores, produced scores on the long-term face memory tests that were below control means, some significantly. This has important implications for policing, as some forces have created specialist SR units whose successes are part based on superior longer-term face recognition ability. If recruitment was based only on the short-term memory tests commonly used in research, the impact on crime detection might be lower than if such a unit contained individuals pre-tested over longer retention intervals. Regardless, tests can only provide a marker of ability. They will never guarantee actual workplace performance in any occupation. Nevertheless, the results of the current research suggest that SRs with exceptional short- and long-term face memory, and simultaneous face matching ability are more likely than controls to identify targets across retention intervals of at least 56 days.

References

Bate, S., Bennetts, R., Hasshim, N., Portch, E., Murray, E. Burns, E., Dudfield, G. (2018). The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. *Journal of Experimental Psychology: Human Perception and Performance, 45*(3), 363-377. doi:10.1037/xhp0000607

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3,* 1-19. DOI: 10.1186/s41235-018-0116-5

Behrmann, M., Avidan, G., Marotta, J. J., & Kimchi, R. (2005). Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioural findings. *Journal of Cognitive Neuroscience, 17,* 1130-1149. doi:10.1162/0898929054475154

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex, 98,* 91-101. doi:10.1016/j.cortex.2018.07.008

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition, 1*(2), 96-103. doi:10.1016/j.jarmac.2012.02.001

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex, 82,* 48-62. doi:10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS one, 11*(2), doi:10.1371/journal.pone.0148148

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2016). Super-Recognizers in action: Evidence from face matching and face memory tasks. *Applied Cognitive Psychology, 30*(1), 81-91. doi:10.1002/acp.3170

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology, 7*, 1378. doi:10.3389/fpsyg.2016.01378

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42,* 286–291. doi:10.3758/BRM.42.1.286.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*, 243-248. doi:10.1111/1467-9280.00144

Clark, S.E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238-259. doi:10.1177/1745691612439584

Courtois, M. R., & Mueller, J. H. (1981). Target and distracter typicality in face recognition. *Journal of Applied Psychology, 66,* 639–645. doi:10.1037/0021-9010.66.5.639

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology, 27,* 245-260. doi:10.1080/02643294.2010.513967

Davis, J. P., Bretfelean, D., Belanova, E., & Thompson, T. (2019, September 5). Super-recognition and long-term memory. DOI:10.17605/OSF.IO/ZMCDH

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*(6), 827-840. doi:10.1002/acp.3260

Davis, J. P., Maigut, A. C., Jolliffe, D., Gibson, S, & Solomon, C. (2015). Holistic facial composite creation and subsequent video line-up eyewitness identification paradigm. *Journal of Visualized Experiments, 106,* e53298. doi:10.3791/53298

Davis, J. P., & Tamonytė, D. (2017). Masters of disguise: Super-recognisers' superior memory for concealed unfamiliar faces. *Proceedings of the 2017 Seventh International Conference on Emerging Security Technologies (EST),* 6-8 September 2017, Canterbury, UK**.** doi:10.1109/EST.2017.8090397

Davis, J. P., Treml, T., Forrest, C., & Jansari, A. (2018). Identification from CCTV: Assessing super-recogniser police ability to spot faces in a crowd, and susceptibility to change blindness. *Applied Cognitive Psychology, 32*(3), 337-353. doi:10.1002/acp.3405

Davis, J. P., & Valentine, T. (2015). Human verification of identity from photographic images. In T. Valentine and J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 211-238). Chichester: Wiley-Blackwell. doi:10.1002/9781118469538.ch9

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior, 30,* 287-307. doi:10.1007/s10979-006-9008-1

Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory

representation. *Journal of Experimental Psychology: Applied, 14*(2), 139. doi:10.1037/1076-898X.14.2.139

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*, 576–585. doi:10.1016/j.neuropsychologia.2005.07.001.

Ebbinghaus, H. (1913). *Memory: A Contribution to Experimental Psychology; Translated by Henry A. Ruger and Clara E. Bussenius*. Teachers College, Columbia University, New York

Esins, J., Schultz, J., Stemper, C., Kennerknecht, I., & Bulthoff, I. (2016). Face perception and test reliabilities in congenital prosopagnosia in seven tests. *i-Perception, 7*(1), 1-37. doi:10.1177/2041669515625797

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education, 74*(3), 249-266. doi: 10.3200/JEXE.74.3.249-266

Grabman, J. H., Dobolyi, D. G., Berlovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: the role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition, 8,* 233- 243. doi:10.1016/j.jarmac.2019.02.002

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Konar, Y., Bennett, P. J., & Sekuler, A. B.(2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science, 21*(1), 38-43. doi: 10.1177/0956797609356508

Lander, K., Christie, F., & Bruce, K. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition, 27*, 974-985. doi:10.3758/BF03201228

Macmillan, N., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Marzi, T., & Viggiano, M. P. (2010). Deep and shallow encoding effects on face recognition: an ERP study. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *78*(3), 239-250. DOI: 10.1016/j.ijpsycho.2010.08.005

Maxwell, S. E., Deaney, H.D., & Kelley, K. (2018). *Designing Experiments and Analyzing Data: A Model Comparison Perspective.* (3rd ed.). New York: Routledge

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, *3*(1), 21. doi:10.1186/s41235-018-0112-9

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory and Cognition, 34*(4), 865-876. doi:10.3758/BF03193433

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law 7*, 3-35. doi:10.1037//1076-8971.7.1.3

Mickes, L., Flowe, H., & Wixted, J. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous vs. sequential lineups. *Journal of Experimental Psychology: Applied, 18*(4), 361-376. doi:10.1037/a0030609

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2017). Face recognition ability does not predict person identification performance: using individual data in the interpretation of group results. *Cognitive Research: Principles and Implications, 3*(23), 1- 13.

doi:10.1186/s41235-018-0117-4

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser? In M. Bindemann & A. M. Megreya (Eds.), *Face processing: Systems, Disorders, and Cultural Differences*. New York, NY: Nova.

O'Keefe, P. (2016, 22 August). The detectives who never forget a face. *New Yorker*. Retrieved from http://www.newyorker.com/magazine/2016/08/22/londons-super-recognizer-police-force

O'Toole, A. J. Roark, D., & Abdi, H. (2002). Recognition of moving faces: A psychological and neural framework. *Trends in Cognitive Sciences, 6*, 261-266. doi:10.1016/S1364-6613(02)01908-3

Police and Criminal Evidence Act (1984). Codes of Practice, Code D. Home Office, Retrieved 10 February 2020 from, https://www.gov.uk/guidance/police-and-criminal-evidence-act-1984-pace-codes-of-practice

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology, 110,* 461-479. doi:10.1111/bjop.12368

Robbins, R.A., & Coltheart, M. (2015). The relative importance of heads, bodies, and movement to person recognition across development. *Journal of Experimental Child Psychology, 138*, 1-14. doi:10.1016/j.jecp.2015.04.006.

Robertson, A. (2016, April 2). Serial thief who pilfered more than £100,000 in jewellery, handbags, clothes and antiques from luxury London boutiques is finally caught after Scotland Yard brings in its team of 'super recognisers. *The Daily Mail.* Retrieved from: https://www.dailymail.co.uk/news/article-3520488/Serial-thief-pilfered-100-000-jewellery-handbags-clothes-antiques-luxury-London-boutiques-finally-caught-Scotland-Yard-brings-team-super-recognisers.html

Robertson, D., Black, J., Chamberlain, B., Megreya, A. M., & Davis, J. P. (2020). Super-recognisers show an advantage for other race face identification *Applied Cognitive Psychology, 34*(1), 205-216. DOI: 10.1002/acp.3608

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police super-recognisers. *PloS One, 11*(2), e0150036–8. doi:10.1371/journal.pone.0150036

Royer, J., Blais, C., Charbonneau, I., Déry, K., Tardif, J., Duchaine, B., … Fiset, D. (2018). Greater reliance on the eye region predicts better face recognition ability. *Cognition, 181*, 12–20. https://doi.org/10.1016/j.cognition.2018.08.004

Russell, R., Duchaine, B., & Nakayama, K., (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*, 252–257. doi:10.3758/PBR.16.2.252

Satchell, L., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2019). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality, 79,* 49-58. doi:10.1016/j.jrp.2019.02.002

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine and J.P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Chichester: Wiley-Blackwell.

Sauer, J. D., & Brewer, N. Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*(4), 337-347. doi:10.1007/s10979-009-9192-x

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12887-12892. doi:10.1073/pnas.1421881112

Shepherd, J. W., & Ellis, H. D. (1973). The effect of attractiveness on recognition memory for faces. *American Journal of Psychology, 86,* 627-633. doi:10.2307/1421948

Shepherd, J. W., Ellis, H. D., & Davies, G. M. (1982). *Identification evidence: A psychological evaluation.* Aberdeen, Scotland: Aberdeen University Press.

Souza, A.S., Rerko, L., & Oberauer. K. (2015). Refreshing memory traces: thinking of an item improves retrieval from visual working memory. *Annals of the New York Academy of Science, 1339,* 20– 31. doi: 10.1111/nyas.12603

Sporer, S. L., Penrod, S. D., Read, D., & Cutler, B. L. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315-327. doi:10.1037/0033-2909.118.3.315

Susilo, T., & Duchaine, B. (2013). Advances in developmental prosopagnosia research. *Current Opinion in Neurobiology, 23,* 423–429. doi:10.1016/j.conb.2012.12.011

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 46*(2), 225-245. doi:10.1080/14640749308401045

Tanaka, J.W., Heptonstall, B., & Campbell, A. (2019). Part and whole face representations in immediate and long-term memory. *Vision Research, 164,* 53 – 61. https://doi.org/10.1016/j.visres.2019.07.007

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE, 14*(2), 1-17. doi:10.1371/journal.pone.0211037

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research, 141*, 217–227. doi:10.1016/j.visres.2016.12.014

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*(8), e103510. doi:10.1371/journal.pone.0103510

Wickham, L. H., Morris, P. E., & Fritz, C. O. (2000). Facial distinctiveness: its measurement, distribution and influence on immediate and delayed recognition. *British Journal of Psychology, 91*(1), 99-123. doi:10.1348/000712600161709

Wiese, H., Komes, J., & Schweinberger, S. R. (2013). Ageing faces in ageing minds: A review on the own-age bias in face recognition. *Visual Cognition, 21*(9-10), 1337-1363. doi: 10.1080/13506285.2013.823139

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, … Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the USA, 107*, 5238e5241. doi:10.1073/pnas.0913053107

Wixted, J. & Mickes, L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications, 3*, 1 – 22. https://doi.org/10.1186/s41235-018-0093-8

Yarmey, A. D. (1979). The effects of attractiveness, feature saliency and liking on memory for faces. In M. Cook and G. Wilson (Eds.), *Love and attraction* (pp. 51–53). Oxford, England: Pergamon Press.

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*, 100-110. doi:10.1016/j.tics.2017.11.007

Tables

Table 1: Between-group tests comparing group constitution as well as CFMT+ and GFMT scores in Experiment 1

	SRs (n = 121)		Controls (n = 165)					
	N	%	*N*	%		χ	*V*	*p*
Gender (male)	45	37.2	81	49.1		4.01	.118	.045
Ethnicity (white)[A]	94	79.7	138	84.1		.946	.058	>.2
	M	*SD*	*M*	*SD*	*df*	*t*	*d*	*p*
Age	33.5	9.9	36.2	13.3	281.9	-1.94	0.23	.065
CFMT+ criteria	≥ 95		58-83					
CFMT+ (max 102)	96.81	1.80	75.10	6.57	196.4	40.42	4.51	<.001
GFMT (max 40)	38.44	1.69	35.81	2.83	273.8	9.79	1.13	<.001

[A] *Note that ethnicity data was missing for 3 SRs and 1 control*

Table 2. Spearman's correlation coefficients between all measures in Experiment 1 (n = 597) [A B]

	GFMT	Retention interval	LTFMT10 Hit Rates	Foil IDs	Misses
CFMT+	0.42 *	0.02	0.23 *	-0.24 *	0.01
GFMT		<0.01	0.20 *	-0.22 *	0.05
LTFMT10					
Retention interval			-0.31 *	0.28 *	0.01
Hit rates				-0.57 *	-0.40 *
Foil IDs					-0.43 *

[A] *Pearson correlations were also conducted and showed similar outcomes*
[B] *All coefficients marked * were significant (p < .001)*

Table 3. Criteria for SR and control groups, and results for t-tests comparing their CFMT+ and GFMT outcomes in Experiment 2

	SRs (n = 301)		Controls (n = 415)			χ^2	*V*	*p*
Gender (male)	78	25.9%	136	32.8%		5.26	.086	.072
Ethnicity (white)	258	85.7%	367	88.4%		1.16	.040	>.2
	M	*SD*	*M*	*SD*	*df*	*t*	*Cohen's d*	*p*
Age	37.84	*10.36*	40.23	*13.31*	709.08	-2.70	0.20	.007
CFMT+ criteria	≥ 95		58-83					
CFMT+ (102)	97.17	*1.78*	75.21	*6.40*	498.78	66.48	4.68	<.001
GFMT (40)	38.57	*1.86*	35.78	*2.66*	712.99	16.53	1.20	<.001

Table 4. Spearman's correlation coefficients across all measures in Experiment 2 (n = 1421) [A B]

					LTFMT20 Outcomes					
	GFMT		Retention interval	TP Hits		TP misses		TP foil IDs		TA CRs	
CFMT+	0.47	*	0.04	0.25	*	-0.02		-0.25	*	0.13	*
GFMT			0.02	0.17	*	-0.02		-0.17	*	0.11	*
LTFMT20											
Retention interval				-0.43	*	0.19	*	0.25	*	-0.04	
TP Hits						-0.45	*	-0.56	*	-0.07	
TP Misses								-0.41	*	0.58	*
TP Foil IDs										-0.48	*

[A] *All coefficients marked * were significant p < .001*

[B] *Pearson correlation analyses generated similar results, except the negative relationship between LTFMT20 hits and CRs became significant, r(1421) = -.08, p = .003.*

Figures

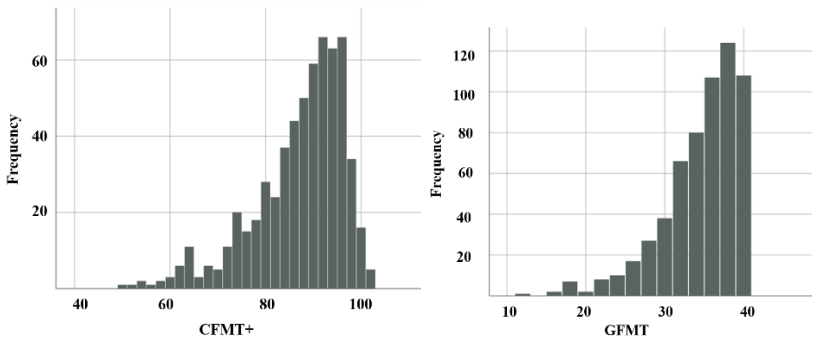

Figure 1. Distribution of CFMT+ and GFMT scores in Experiment 1 (n = 597)

Figure 2. Still images of the 10 actor targets displayed in Phase 1 (originally colour) videos

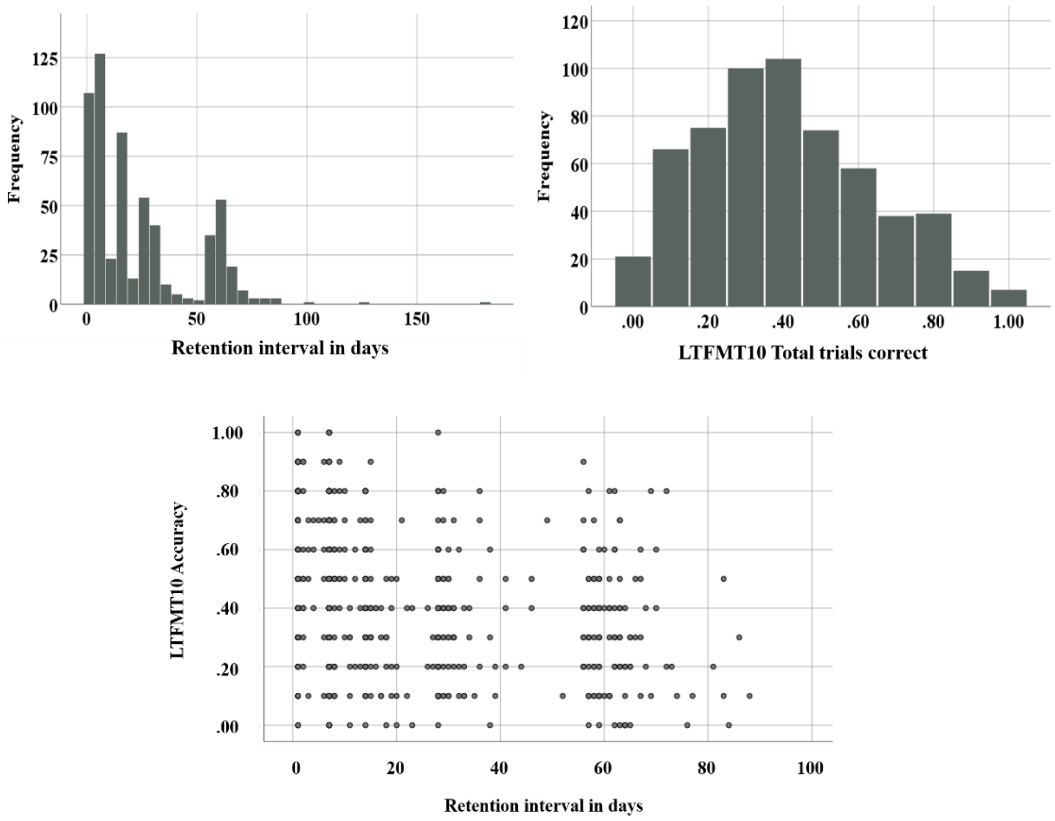

Figure 3. For all participants in Experiment 1 (n = 597), top left: LTFMT10 retention interval, top right: hit rates, and bottom: a scatterplot displaying both (for clarity three participants with outlier retention intervals (> 100 days) were excluded, n = 594).

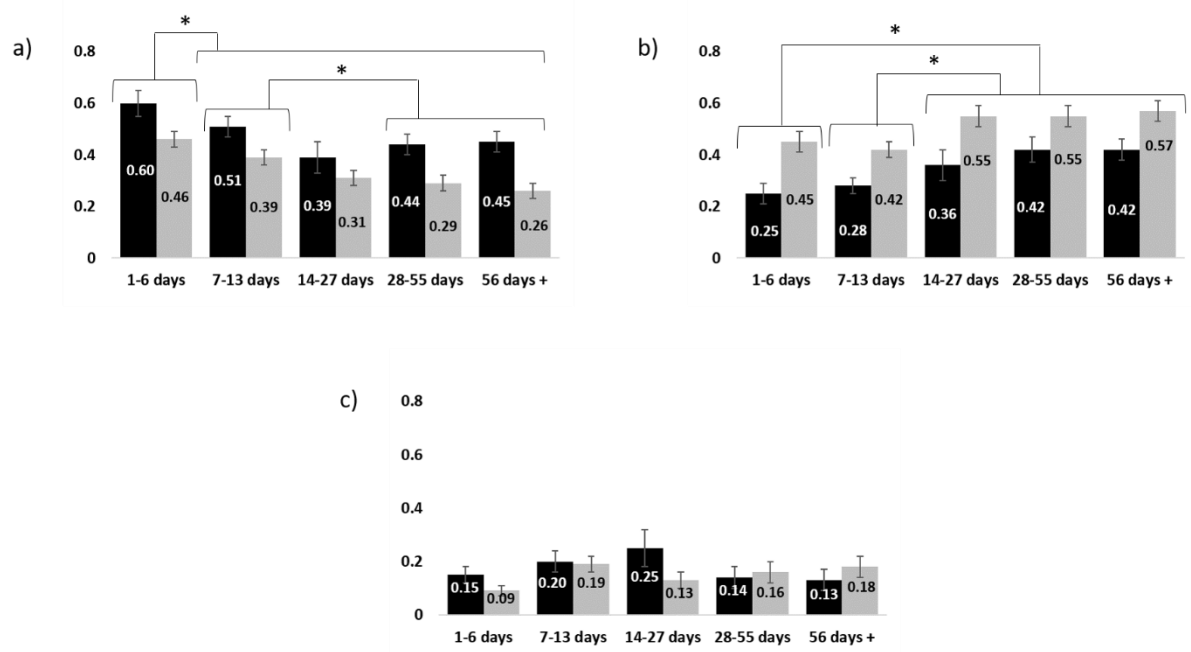

Figure 4: LTFMT10 (a) hits (b) foil IDs, and (c) misses in Experiment 1 as a function of group and retention interval (dark bars = SRs, grey bars = controls)(error bars = 1 SEM). Significant values indicate Tukey post-hoc test differences.

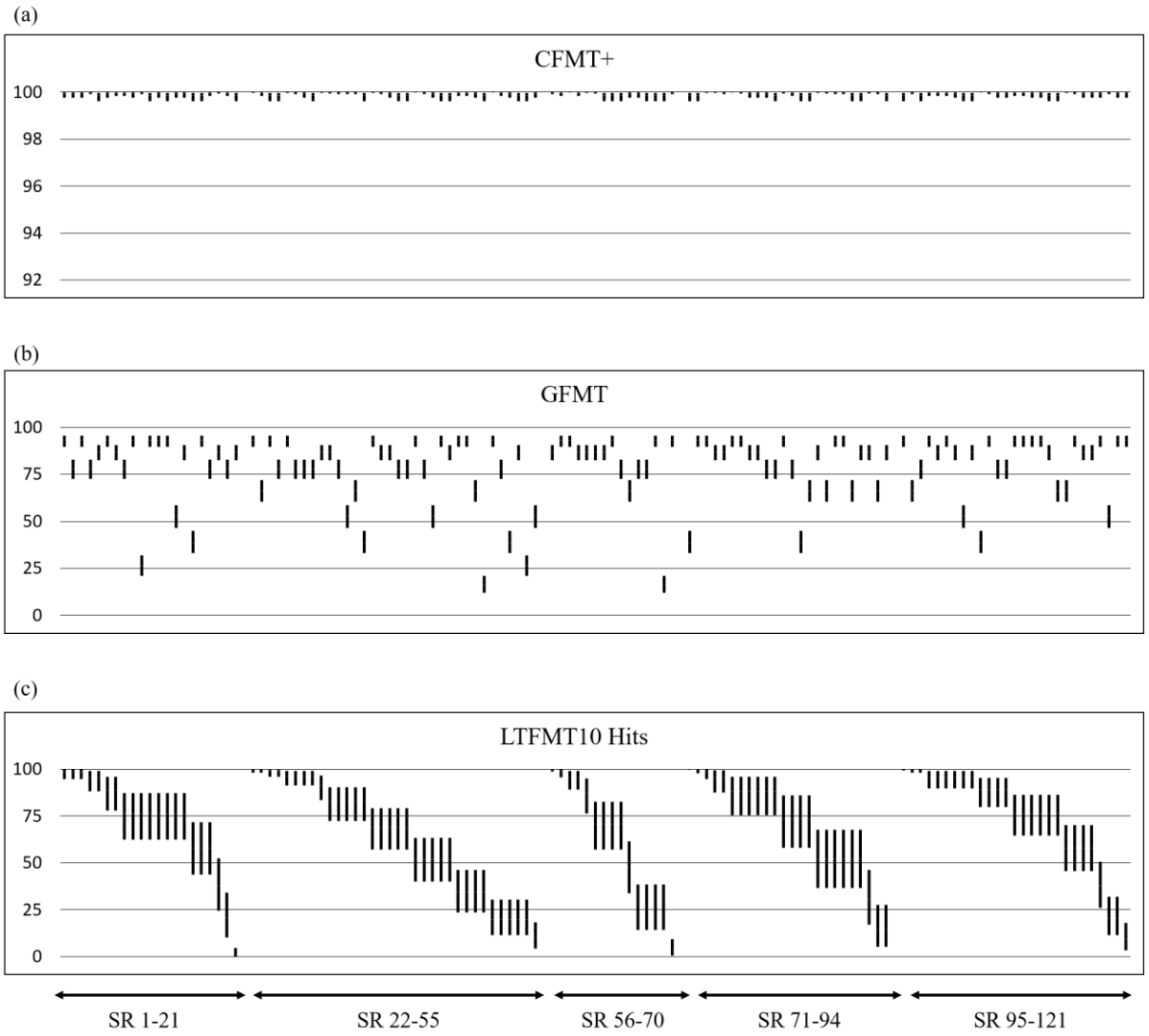

Figure 5.Upper and lower bound confidence intervals (95%) of the estimated proportion of the population expected to fall below each super-recogniser (SR) (n = 121) based on (1a) CFMT+ scores, (1b) GFMT, and (1c) LTFMT10 hits. Due to low super-recogniser variability on the CFMT+ only 92%-100% range is depicted in Figure 1a. The 50% line on Figures 1b and 1c represents the control mean, so that 50% of the population would be expected to achieve above this level. To enhance interpretability throughout, super-recognisers are grouped based on delay condition (1-6, 7-13, 14-27, 28-55, 56+ days), and rank-ordered from left-to-right based on LTFMT10 hit rates (Experiment 1).

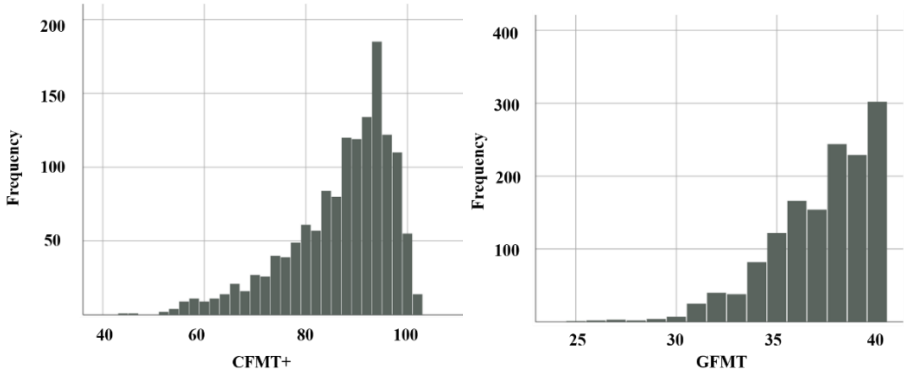

Figure 6. Distribution of CFMT+ and GFMT scores in Experiment 2 (n = 1421)

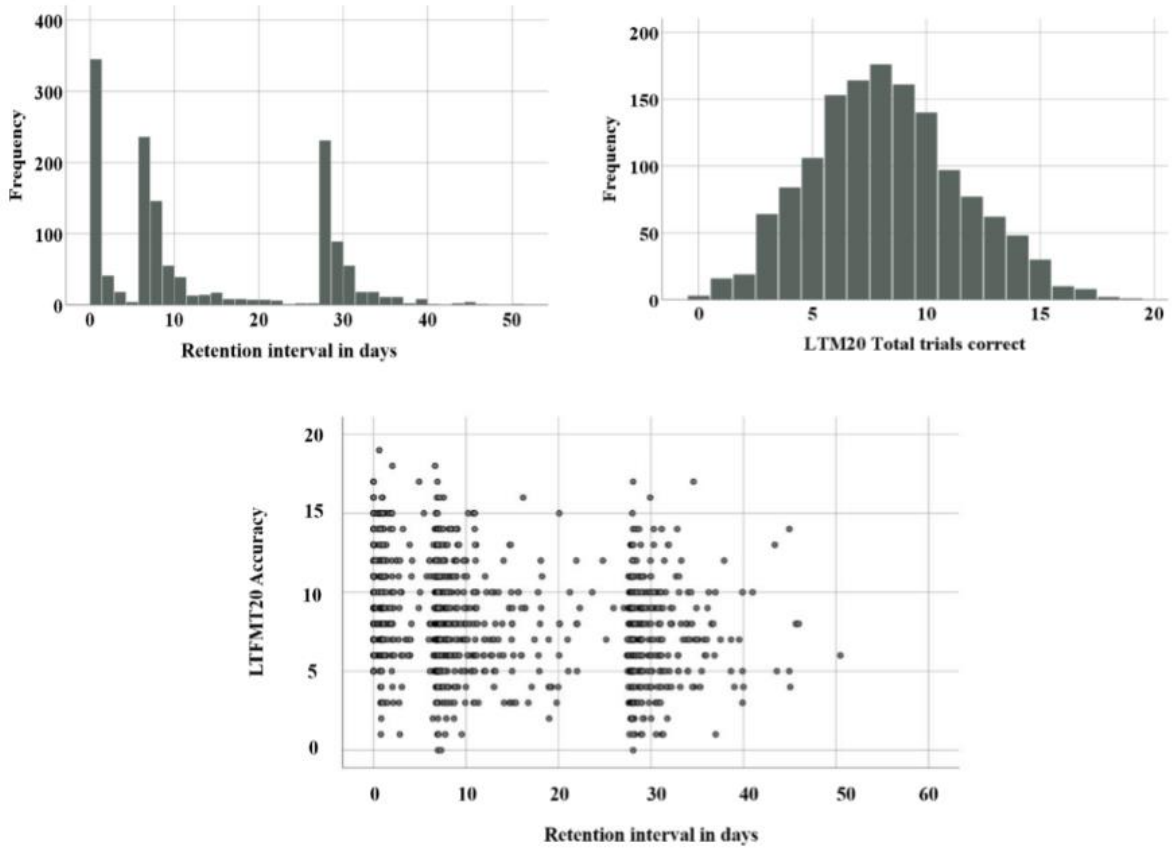

Figure 7. Retention interval in days, total correct responses regardless of target-presence, and a scatterplot displaying the proportion of correct responses (max = 20) and retention interval on the LTFMT20 in Experiment 2 (n = 1,421)

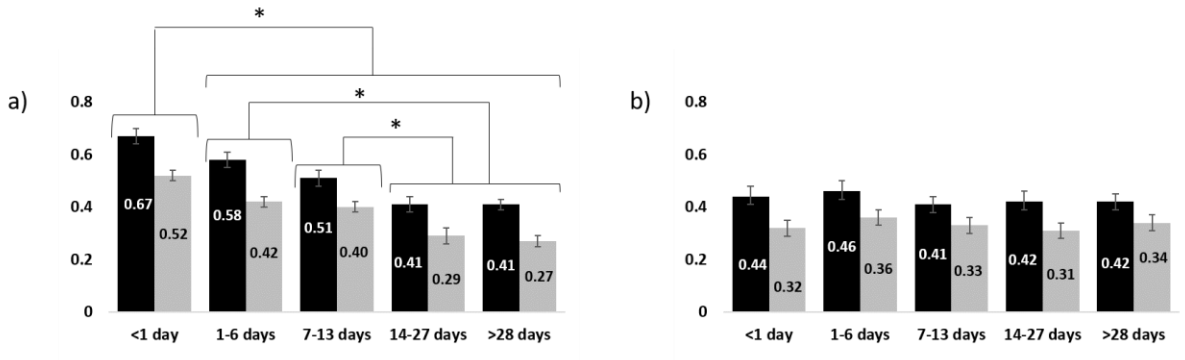

Figure 8. a. Hits and b. CRs by SRs and controls in each retention interval condition (< 1 day: SRs: n = 53, controls n = 92; 1-6 days: SR: n = 56, controls: n = 91; 7-13 days: SRs: n = 66, controls: n = 87; 14-27 days: SRs: n = 41, controls: n = 59; ≥ 28 days: SRs: n = 85, controls n = 86). (dark bars = SRs, grey bars = controls) (error bars = 1 SEM). Significant values indicate Tukey post-hoc test differences.

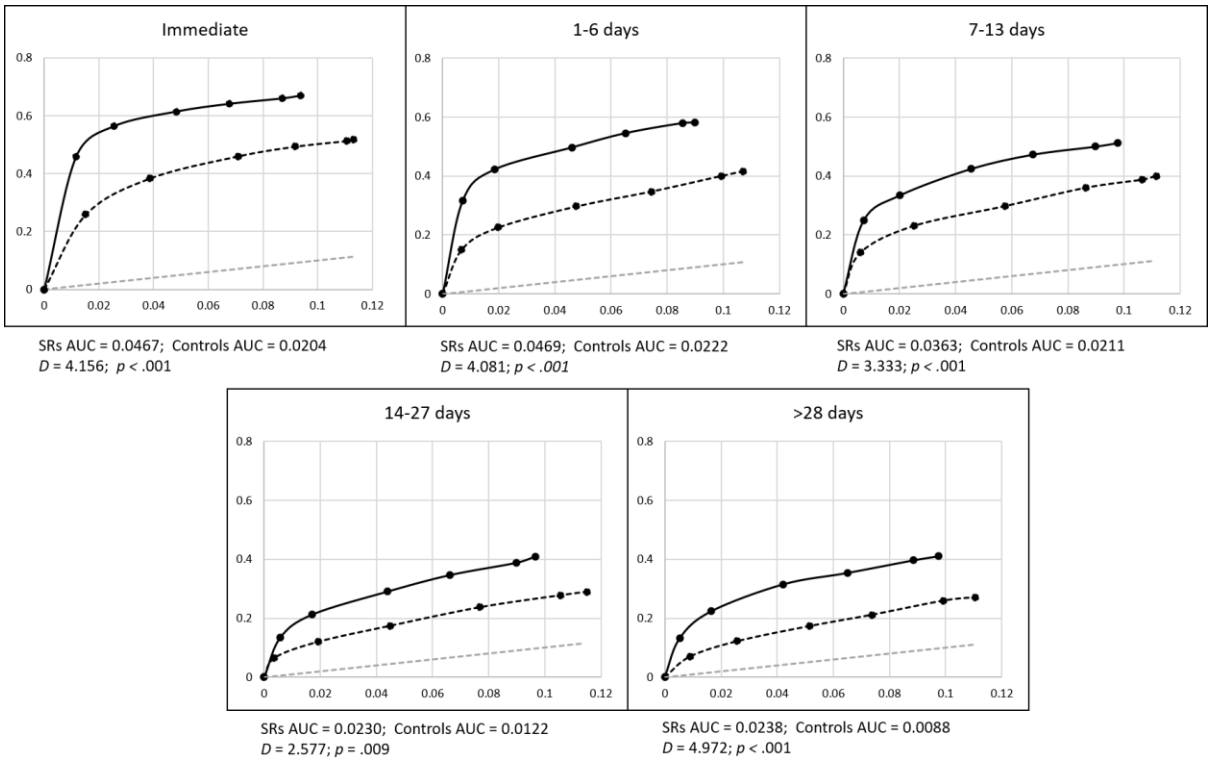

Figure 9. ROC analysis for SRs (uninterrupted line) and controls (dashed line) in the five retention intervals with confidence-based accuracy (hits) plotted as a function of confidence-based false alarms (FA) across all trials. The vertical axis depicts the hit rate, the horizontal axis the FA rate. The grey line indicates performances at chance levels (i.e. when hit rates equal false positive rates). AUC are indicated for SRs and controls in each retention interval condition. The points on each line represent diagnosticity ratio for each confidence range. From left to right: >90% confidence; >70%; >50%; >30%;>10%;>1% .

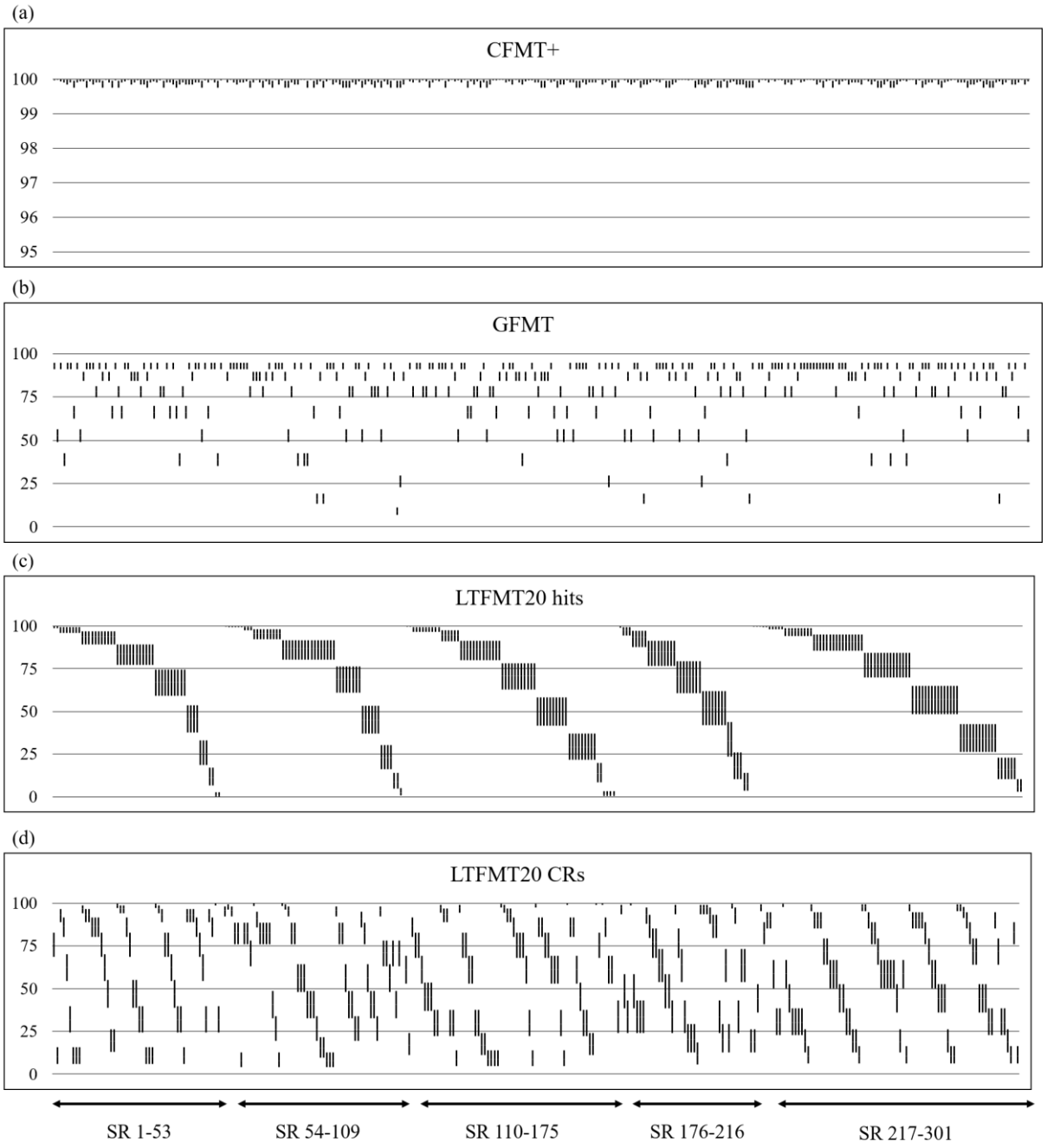

Figure 10.Upper and lower bound confidence intervals (95%) of the estimated proportion of the population expected to fall below each super-recogniser (SR) (n = 301) based on (2a) CFMT+ scores, (2b) GFMT, (2c) LTFMT20 hits, and (2d) LTFMT20 CRs. Due to low super-recogniser variability on the CFMT+ only 95%-100% range is depicted in Figure 2a. The 50% line on Figures 2b, 2c, and 2d represents the control mean, so that 50% of the population would be expected to achieve above this level. To enhance interpretability throughout, SRs are grouped based on delay condition (immediate, 1-6, 7-13, 14-27, 28+ days), and rank-ordered from left-to-right based on LTFMT20 hit rates (Experiment 2).