Assessing the long-term face memory of highly superior and typical-ability short-term face recognisers

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

Department of Psychology, Social Work and Counselling, 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

Dr Trevor Thompson, Senior Lecturer in Psychology, London, SE10 9LS, t.thompson@gre.ac.uk; +44(0)208 331 9632

L. Diandra Bretfelean, Research Associate, University of Greenwich, London, SE10 9LS, luiza.bretfelean.17@ucl.ac.uk

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 for each of the actors. They would also like to thank Monika Durova and Lauren Jensen for assistance with data collection. Parts of these data were presented at the British Psychological Society: Cognitive Section Annual Conference, Newcastle UK, 30 August 2017, as a poster by the second author after winning the section's 2017 Annual Undergraduate Project Prize (Bretfelean & Davis, 2017). This project received ethical approval from the University of Greenwich Research Ethics Committee.

Word Count: 11,793

Abstract

Exceptional long-term face recognition ability is the hallmark of super-recognition. However, previous super-recognition research has employed brief retention intervals, and therefore using eight, or ten-target eyewitness identification designs, this was the first to demonstrate that many super-recognisers' maintain unfamiliar face recognition superiority after long delays. In Experiment 1, with delays of at least seven days, compared to controls (*n* = 222), longer Phase 1 target-video exposure, and novel hybrid-video array-style Phase 2 line-ups, disproportionally facilitated super-recognisers' (*n* = 112) correct identifications, and rejections of previously unseen faces. Standardised line-up target presence instruction impact also differed between-groups. In Experiment 2, super-recognisers (*n* = 57) displayed stronger commitment effects than controls (*n* = 103) to repeat their correct and incorrect identification decisions in a second line-up despite delays of 36-275 days (*M* = 171). In Experiment 3, super-recognisers' (*n* = 57) superiority over controls (*n* = 103) after one day was at the same level as after 56 days implying ability-independent facial memory decay. Throughout, between-group effect sizes were stronger than those of delay, while there were substantial individual within-group differences, suggestive of super-recogniser sub-types. Worldwide, police deployment of super-recognisers has positively impacted crime detection, and with the inclusion of all participants (*n* = 1,688), suggestions were made as to minimum super-recogniser employment recruitment criteria. This first systematic investigation into their long-term face memory suggests that face recognition tests with substantial retention intervals should be included to ensure possession of a full complement of abilities.

Keywords

Face recognition, Long-term memory, super-recognition, line-up, identification

Introduction

Substantial individual differences in face recognition ability in the neuro-typical population range from *developmental prosopagnosics* who struggle to recognise familiar faces (e.g., Rossion *et al.,* 2003), to *super-recognisers* with exceptional skills with unfamiliar faces (e.g., Russell, Duchaine, & Nakayama, 2009; for a review see Noyes, Philips, & O'Toole, 2017). Differences are mainly face-specific and inherited (e.g. Shakeshaft & Plomin, 2015; Wilmer et al., 2010), but impacted by exposure (e.g. cross-age effects, Belanova, Davis, & Thompson, 2018; cross-ethnicity effects, Meissner & Brigham, 2001). Extensive research has investigated developmental prosopagnosia, while recently a growing body of research has examined super-recognisers' superiority at unfamiliar short-term face memory (e.g. Bate et al., 2018; Russell et al., 2009), simultaneous face matching (e.g., Bobak, Dowsett, & Bate, 2016; Bobak, Parris, Gregory, Bennetts, & Bate, 2016), and long-term *familiar* face memory (e.g., Davis, Lander, Evans & Jansari, 2016; Russell et al., 2009). Anecdotally, super-recognisers often describe 'feats' of exceptional long-term unfamiliar face memory (e.g. Russell et al., 2009), and yet only one published report has investigated this using delays of a week (Davis & Tamoytė, 2017). Employing only one target-actor however, generalisability was limited. The current research aimed to address this gap in the literature.

This research has important theoretical and practical implications. Investigating super-recognisers may enhance theoretical understanding of the nature of individual differences in unfamiliar face processing ability (Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Bobak, Parris et al., 2017). In addition, some police forces employ super-recognisers in roles utilising their skills (Davis et al., 2016; Davis, Lander, & Jansari, 2013; Davis, Treml, Forrest, & Jansari, 2018; Edmond & Wortley, 2016: Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). Understanding super-recognisers' aptitudes and limitations may assist management to make informed deployment decisions, as well as to assist in recruitment test development.

Definition of super-recognition

Super-recognisers outperform controls on a range of face processing tasks (Bobak, Dowsett et al., 2016; Bobak, Parris et al., 2016; Davis et al., 2016; 2018), implicating quantitative differences in ability, although there is also growing evidence of qualitatively different neurological and cognitive mechanisms (e.g., Belanova et al., 2018; Bobak, Bennetts et al., 2016; Russell et al., 2009). Some super-recognisers and developmental prosopagnosics display substantial between-test and between-domain (memorial/perceptual) score variability, suggesting these constructs might be heterogeneous in nature (e.g. Bate & Tree, 2017; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock et al., 2016). Nevertheless, there is currently no agreed scientific definition or performance threshold for super-recognition (Noyes et al., 2017). Inclusion criteria for extreme-ability groups in research have mainly been based on self-beliefs in poor (developmental prosopagnosia), or outstanding (super-recogniser) abilities, although many self-professed super-recognisers vastly over-estimate their true ability (e.g. Bate et al., 2018). Beliefs are normally supported by low or high performances respectively on face processing tests (see Noyes et al., 2017). Using statistical definitions, group allocation has commonly been based on scores at least 2 standard deviations (SD) below (developmental prosopagnosia) or above (super-recognition) the estimated population mean on short-term unfamiliar face recognition tests (i.e. bottom and top 2%). The standardised *Cambridge Face Memory Tests* (CFMT, Duchaine & Nakayama, 2006; *Extended:* CFMT+, Russell et al., 2009) have often been employed, although there has been some criticism of their utility (e.g., Bate et al., 2018; Esins, Schultz, Stemper, Kennerknecht, & Bulthoff, 2016). Nevertheless, even though statistically-based exceptionally good performances on empirical tests may be highly indicative, some individuals achieving very high scores may not in fact be super-recognisers. A series of ‘lucky’ guesses or deductions on multiple-choice tests (e.g. CFMT+) may inflate performance. Likewise, poor scores by genuine super-recognisers may be due to factors like fatigue or distractions.

Allocation to super-recogniser groups is likely to be more reliable if based on multiple test scores conducted in the controlled confines of a laboratory or an examination (Noyes et al., 2017). As the research described in this paper was entirely internet-based, the term *superior-face-recogniser* (SFR) is employed for participants who met previous super-recognition research-based criteria. Pre-test self-assessments of ability were collected. However, due to their unreliability, superior-face recogniser and control group inclusion criteria were based only on rigorous CFMT+ thresholds (Bobak, Pampoulov, & Bate, 2016); as well as an ability verifying unfamiliar *Short-Term Face Memory Test* (STFMT).

Long-term face memory

Some research with participants scoring in the typical range has examined face recognition accuracy using retention intervals of 4 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). The forgetting curve for the human face embraces a form determined by Ebbinghaus (1913) for other stimuli. Although factors such as facial distinctiveness (Wickham, Morris, & Fritz, 2000), initial memory strength (Deffenbacher et al., 2008), and repeated procedures (e.g. Deffenbacher, Bornstein, & Penrod, 2006) have an impact; most forgetting occurs in the first 24 hours, and gradually increases over longer periods (Deffenbacher et al., 2008). If SFRs' ability is driven by enhanced face *encoding*, then their forgetting curve might be predicted to match the shape of controls. However, if superiority is driven by enhanced long-term memory, then a shallower forgetting curve than controls might be expected.

The primary aim of the three experiments described here was to compare the performance of SFRs with controls on tests of long-term face memory. In Experiment 1, we used delays of at least a week and varied factors known to impact eyewitness line-up identification performance. In Experiment 2, the same participants re-viewed the same line-ups after mean delays of about 6 months. In Experiment 3, with new participants, we employed only target-present line-ups with delay varied from one-day to over 56-days. Previous research found positive short- and long-term face memory relationships (e.g. Bindemann, Brown, Koyas, & Russ, 2012); with longer delays resulting in lower accuracy (e.g. Deffenbacher et al., 2008). Similar effects were predicted here.

Experiment 1

In Experiment 1, SFRs and 'average-ability' controls who had previously completed the CFMT+ (Russell et al., 2009); were assessed on the ability-verifying STFMT, and an 8-trial *Long-Term Face Memory Test* (LTFMT8). Longer exposures promote greater encoding opportunities facilitating recognition (e.g., Memon, Hope, & Bull, 2003; for meta-analyses, see Bornstein, Deffenbacher, Penrod, & McGorty, 2012; Shapiro & Penrod, 1986), and to measure this in Phase 1 of the LTFMT8, participants viewed four 30s and four 60s target-actor videos. At least seven days later they viewed four counterbalanced target-present and four target-absent photo or *hybrid-video* simultaneous line-ups. In the latter, videos were sequentially played while participants simultaneously viewed the entire array. Approximately half were *warned* that the line-ups 'may or may not contain the targets', instructions provided by most police jurisdictions to reduce innocent suspect identifications by eyewitnesses. However, the warning reduces incorrect *and* correct line-up selections (e.g. Clark, 2005).

Context dependent memory theories propose facilitated performance when learning and test conditions match (e.g. Tulving & Thomson, 1973), suggesting that compared to photo line-ups, accuracy should be higher with video line-ups matching Phase 1 moving displays. Line-up research supports this (e.g., Havard, Memon, Clifford, & Gabbert, 2010; for a review see Fitzgerald, Price, & Valentine, 2018), and movement can enhance unfamiliar face recognition (e.g. O'Toole, Roark, & Abdi, 2002), which may partly be due to additional views of targets in multiple video frames providing identification cues, rather than movement itself (although see Lander, Christie, & Bruce, 1999). On the other hand, movement reliably facilitates familiar face recognition in poor-quality images (Lander & Chuang, 2005; Lander & Davies, 2007), and becomes a more important identity cue with greater familiarisation (O'Toole et al., 2002). If their superior skills are partly due to enhanced face familiarisation, SFRs might be expected to show a pattern of results associated with familiar faces and gain a disproportionate unfamiliar face recognition advantage over controls from movement in hybrid-video line-ups.

Therefore, for all participants in Experiment 1, accuracy was predicted to be enhanced by longer Phase 1 displays, while the biasing effect of the target-presence warning was expected to reduce line-up selections. Bate et al. (2018) suggests that super-recognisers' superiority in tests may be driven by recognising that a face has not been seen before, as much as recognition of previously seen faces. Therefore, compared to controls, SFRs were expected to make more target-present line-up correct identifications (hits), and target-absent line-up correct rejections (CRs). They were also expected to gain a disproportionate advantage from the hybrid-video over photo line-ups.

Method

Design

In the LTFMT8 Phase 1, participants viewed eight target-actor videos. In Phase 2 after at least a week, they identified the actors from line-ups. In a five-way design, between-subjects factors were *group* (SFRs *vs*. controls), with membership based on CFMT+ (Russell et al., 2009), and STFMT scores; *line-up media* (hybrid-video *vs.* photo); and *warning* (warning vs. no warning), that targets 'may or may not be depicted in the line-ups'. Within-subjects factors were Phase 1 *exposure time* (30s *vs.* 60s), and line-up *target-presence* (target-present *vs.* target-absent). The dependent variables were target-present line-up hits, target-absent line-up CRs; and signal detection theory (SDT) measures of sensitivity [$d^{/}$= *z*(False Alarms = 1 – CRs) – *z*(Hits)], and response bias (C: Criterion = -0.5[*z*(Hits) + *z*(False Alarms)] (Green & Swets, 1966; Macmillan & Creelman, 2004). Delay between phases and confidence (0%: guessing to 100%: absolutely certain) were assessed. Face recognition ability self-beliefs (5 = well above average to 1 = well below average) were collected prior to the STFMT. To examine which factors best predicted LTFMT outcomes, a correlational design with data from all experiments is included in the General Discussion.

Materials

Short-Term Face Memory Test (STFMT): This test's learning phase sequentially displayed 30 frontal-view colour photos of young adult white-Caucasian males wearing identical shirts for 10-sec (kindly provided by Meissner, Brigham, & Butz, 2005). In the immediate test phase, 60 males in varied clothing were displayed. No hairstyle was cropped. With no time limits, participants judged whether faces were 'old' (30 faces) or 'new' (30 faces). Hits (*n* = 1,091) ranged from 0.33-1.00 (*M* = 0.82); CRs from 0.40 to 1.00 (*M* = 0.82, *SD* = 0.11). These were converted to SDT $d^{/}$ and C measures.

Long-Term Unfamiliar Face Memory Test 8 (LTFMT8): Phase 1 consisted of two versions. In each, eight counterbalanced sequentially-presented colour 30s (*n* = 4) and 60s (*n* = 4) videos all displaying some close-up facial views depicted unfamiliar white-Caucasian target males (*n* = 4) and females (*n* = 4) ("Actors A-H") in good outdoor or indoor lighting. Clothing and actions differed to assist discrimination. For instance, some actors walked towards the camera, others sat or stood behind a table. One played golf. No actor possessed distinguishing facial marks, or other foci of attention. For Phase 2, actors wore different clothing, and PROMAT™ video line-ups (Promat Envision International, Nelson, Lancashire, UK) were created at a London police station by an experienced Metropolitan

Police Service officer. A 15-sec head-and-shoulders video was filmed of each actor, and the officer selected nine foils of the same age, ethnicity and 'position in life' from a database of over 23,000 videos (see Davis, Maigut, Jolliffe, Gibson, & Solomon, 2015, for a video depicting the PROMAT system).

Pilot participants (*n* = 40) viewed the eight Phase 1 videos, and seven days later, eight 3 x 3 target-present arrays of stills from each PROMAT-video. Performance at actor identification was near floor and to make the test easier, the three foils most often incorrectly selected were excluded from line-ups, as these were probably the hardest to distinguish from targets. The final line-ups were 3 x 2 arrays 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. One of the three excluded foils described above replaced the actor in target-absent line-ups.

Photo line-ups: Simultaneous 3 x 2 line-ups consisted of two colour stills (frontal view, right profile) of each actor and the five foils.

Hybrid-video line-ups: Normally PROMAT line-ups are displayed sequentially. However, simultaneous line-ups may generate higher accuracy (e.g. Mickes, Flowe, & Wixted, 2012, although see Wells, Smalarz, & Smith, 2015). Therefore, a 3 x 2 *hybrid-video* design was created, so that from top left, to bottom right, each line-up member's video played sequentially (12s each), while the remaining still images were visible. The sequence repeated until a response was made. To reduce display time, 1.5s was cut from the start and end of the 15s PROMAT videos. This did not introduce anomalies as foils and targets faced forward.

Participants

Invited uncompensated participants (*n* = 6,787) had completed the CFMT+ (Russell et al., 2009) in previous unpublished research. This 102-trial test requires identification of six white-Caucasian male hairstyle-cropped targets, from increasingly degraded images. With a representative sample (*n* = 254, *M* = 70.72, *SD* = 12.32), Bobak, Pampoulov et al. (2016) suggest 95/102 best represents a score 2 SD above the mean of the estimated population, and this was minimum SFR criteria here, achieved by *n* = 152).

After removing duplicate entries, 1,388 started the STFMT and 1,091 completed LTFMT8 Phase 2. Indicative of a bias towards recruiting good face recognisers, final participants (*n* = 1,091; male = 426, female = 662; white-Caucasian = 942 (86.5%), aged 18-72 years, *M* = 35.0, *SD* = 11.1) [1] outperformed by more than 1 SD, Bobak, Pampoulov et al.'s (2016) CFMT+ mean (*M* = 84.9, *SD* = 10.0 *vs. M* = 70.7, *SD* = 12.3), *t*(1090) = 46.55, *p* < .001, *d* = 1.26.

To reduce the recruitment bias influence, provisional criterion for "average" face recognition ability controls was a score within 1 SD of Bobak, Pampoulov, et al.'s (2016) mean (i.e. CFMT+: 58-83). The STFMT ($d^{/}$) verified ability (*n* = 370) ($d^{/}$ range = 0.34-3.62; *M* = 1.74, *SD* = 0.58). Those with scores more than 1 SD below the control mean (e.g. $d^{/} \geq 1.16$), suggesting 'worse than average ability', were excluded (*n* = 61).

The same strategy could not be used to exclude 'higher than average ability' controls as many of their STFMT $d^{/}$ scores overlapped with those of provisionally-eligible SFRs (*n* = 152, $d^{/}$ = 0.96-3.67; *M* = 2.43, *SD* = 0.57). Therefore, to ensure all SFRs outperformed all controls, STFMT threshold was the mid-point ($d^{/}$ = 2.085) between SFR and control means, excluding 40 SFRs and another 85 controls (total = 148). Table 1 depicts final group inclusion criteria, mean ability self-belief, CFMT+, and STFMT scores. Independent-measures t-tests demonstrated that with strong effect sizes, SFRs provided significantly higher self-ratings, and outperformed controls on accuracy outcomes, while controls displayed a weak STFMT conservative response bias to be more likely to respond 'new'.

Procedure

[1] Some demographic data were missing.

Invitees were e-mailed a Qualtrics link (www.qualtrics.com) and warned not to use tablets/mobiles to optimise image size. After providing informed consent, which included a request to access previous CFMT+ data; demographic and face memory self-belief data were collected.[2] Participants then completed the STFMT, before starting LTFMT8 Phase 1 which commenced with a cartoon character practice trial and an immediate Phase 2 style target-present line-up. Participants then viewed the eight target-actor (A-H) Phase 1 videos (4 x 30s, 4 x 60s). Passwords and time-limited displays ensured no replays. However, after each video, a question checked whether it had played properly. If participants clicked ‘no’, that video was repeated. In total, 159 (14.6%) participants reported problems, with 221 of the 8,728 videos (2.5%) repeating. This had no effect on any reported results (*p* > .2).

Table 1: Criteria for SFR and control groups, and results for t-tests comparing their self-belief in ability (1-5), CFMT+, and STFMT outcomes in Experiment 1

	SFRs		Controls		*df*	*t*	*d*	*p*
n	112		222					
	SFR and control group inclusion criteria							
CFMT+	95-102		58-83					
STFMT d$^{/}$	> 2.085		1.16-2.085					
Age	18-63 years		18-69 years					
	M	*SD*	*M*	*SD*				
	34.3	8.6	35.4	11.9	292.64	<1	0.11	>.2
Gender								
Male	54 (48.6%)		93 (42.3%)			[A]		
Ethnicity								
White	86.6%		89.6%			[B]		
	M	*SD*	*M*	*SD*				
Self-belief	4.31	0.75	3.72	0.77	332	6.74	0.78	<.001
CFMT+	96.90	1.89	74.81	6.56	284.55	46.48	4.58	<.001
STFMT								
Hits	0.91	0.07	0.78	0.09	272.92	13.80	1.61	<.001
CRs	0.88	0.07	0.79	0.09	285.98	10.54	1.12	<.001
d$^{/}$	2.68	0.41	1.65	0.26	157.03	23.91	3.00	<.001
C	-0.07	0.33	0.01	0.29	332	-2.42	0.26	.016

[A] Gender (male = 1 *vs*. female = 0) x group, χ^2 (1, 331) = 1.21, *p* > .2. Missing data (*n* = 3)
[B] Ethnicity (white = 1 *vs*. other ethnicity = 0) x group, χ^2 (1, 333) < 1. Missing data (*n* = 1).

A week after Phase 1, all participants automatically received Phase 2 URL link e-mails, although not all took part immediately. They were randomly assigned to a line-up

[2] Note: The e-mail also requested use of past *Glasgow Face Matching Test* (GFMT) data (Burton, White, & McNeill, 2010). These data are reported in the General Discussion.

media, and warning condition and identified actors by selecting a line-up number (1-6), or they rejected the line-up ('none of the above'). They were finally debriefed.

Results

LTFMT8 delay was skewed (Shapiro-Wilk (1091) = 0.308, *p* < .001) and varied from 7-87 days, although most participants (87.1%) finished Phase 2 within 10 days of Phase 1 (Median = 7.4; *M* = 9.0, *SD* = 6.0). However, SFR and control delay did not differ, *t*(332) = 1.28, *p* = .200; and the results of delay and counterbalanced within-condition analyses are not reported, as although there were differences in performances indicative of task difficulty variability, there was no interaction on any between-group outcome reported below (*p* > .2). Two (out of 1,091) participants achieved LTFMT8 maximum of 8 out of 8 (*n* = 21 scored 7, *n* = 34 scored 0) (Median = 3, *M* = 3.06, *SD* = 1.59). The full participant sample results (*n* = 1091) are reported in the general discussion. Here only SFR and control data are analysed. Unless otherwise reported, to protect against Type-I errors α = 0.05, and the Bonferroni correction was applied to post-hoc analyses.

Table S1 (supplementary files) depicts mean group outcomes. These were analysed by seven 2 (group: superior-face-recogniser (SFR), control) x 2 (Phase 1 display time) x 2 (line-up media) x 2 (warning) mixed ANOVAs (see Table S2). Group main effects were significant for hits [SFRs (*M* = 0.47, *SD* = 0.29) *vs*. controls (*M* = 0.36, *SD* = 0.25)]; foil IDs [controls (*M* = 0.38, *SD* = 0.29) *vs.* SFRs (*M* = 0.31, *SD* = 0.29)]; CRs [SFR (*M* = 0.45, *SD* = 0.32) *vs*. controls (*M* = 0.32, *SD* = 0.27)]; sensitivity ($d^{/}$) [SFRs (*M* = < .01, *SD* = 1.73) *vs.* controls (*M* = -0.87, *SD* = 1.32)]; and confidence [SFRs (*M* = 62.4, *SD* = 19.7) *vs.* controls (*M* = 53.3, *SD* = 20.1)]. SFRs outperformed controls; and were more confident.

The Phase 1 display time main effects were significant for misses and confidence. Miss rates were higher after 30s (*M* = 0.30, *SD* = 0.33) than 60s (*M* = 0.21, *SD* = 0.29) displays. Confidence was higher after 60s (*M* = 57.2, *SD* = 21.7) than 30s displays (*M* = 55.5, *SD* = 21.2).

The Phase 2 line-up media main effects were significant for hits. Hybrid-video line-ups (*M* = 0.42, *SD* = 0.27) generated more hits than photo line-ups (*M* = 0.35, *SD* = 0.27).

Significant warning main effects for hits [warning (*M* = 0.36, *SD* = 0.26) *vs*. no warning (*M* = 0.43, *SD* = 0.27)]; misses [warning (*M* = 0.30, *SD* = 0.24) *vs.* no warning (*M* = 0.19, *SD* = 0.22)], CRs [warning (*M* = 0.42, *SD* = 0.28) *vs*. no warning (*M* = 0.30, *SD* = 0.29); and criterion [warning (*M* = -0.12, *SD* = 0.64) *vs*. no warning (*M* = -0.46, *SD* = 0.74)] show it induced a conservative response bias, reducing hits, but increasing misses and CRs.

The group x warning interaction was significant. Simple effects of warning found that the warning had no impact on SFR confidence. However, control confidence was higher following the warning than after no warning.

The display time x line-up media interaction was significant for hits. Post-hoc t-tests found no between-line-up media difference with 30s Phase 1 display times. With 60s Phase 1 displays, there was a significant advantage for hybrid-video over photo line-ups.

The stimuli-type x warning interaction was significant for hits. The warning did not influence hybrid-video line-ups, but it significantly reduced photo line-up hit rates.

The group x line-up media x warning interaction was significant for hits. Simple interaction effects by group found that for SFRs, the two-way interaction was significant only, *F*(1, 108) = 5.71, *p* < .05, η^2 = .050. The warning reduced photo line-up hit rates but not video line-ups. No effects were significant for controls (*p* > .1).

The four-way interaction was significant for hits. Simple three-way interaction effects found that for SFRs, only the two-way interaction between line-up media and warning was significant in the direction described above, *F*(1, 108) = 6.79, *p* < .05, η^2 = .059. For controls, only the three-way interaction was significant, *F*(1, 218) = 5.39, *p* < .05, η^2 = .024. With controls only, two display time x line-up media ANOVAs revealed no effects in the warning condition ($p > .2$), although there was a significant interaction after no warning, $F(1, 105) = 6.32$, $p < .05$, $\eta^2 = .057$. Post-hoc tests were not significant ($p < .1$), except there was a marginal possibly spurious effect for photo line-up hit rates to be slightly lower after Phase 1 displays of 60s than 30s ($p < .1$).

The criterion four-way interaction was also significant. Post-hoc analyses however on each simple three-way interaction combination found no significant effects or interactions ($p > 1$), apart from the main effect of warning described above.

Individual analyses: Due to condition numbers, individual analyses comparing each SFR's performance against control means was not feasible. However, Figure 1 displays SFR and control LTFMT8 scores out of 8 regardless of condition or delay demonstrating the overlap between groups. Some short-term SFRs performed very poorly on the LTFMT8, whereas a small proportion of controls outperformed most SFRs, although some effects were likely driven by the between-counterbalanced condition variations in task difficulty.

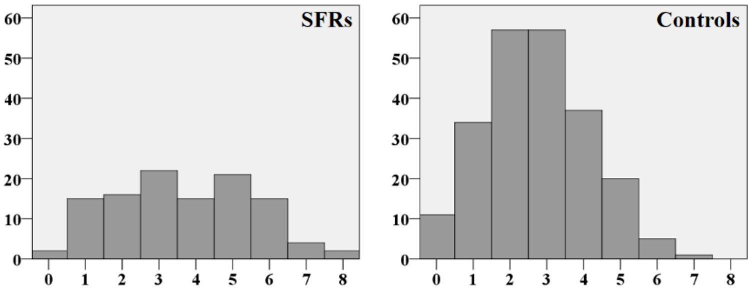

Figure 1: Frequency of individual performances on the LTFMT8 regardless of condition or delay by superior-face-recognisers (SFRs) and controls.

Discussion

Consistent with expectations, Experiment 1 revealed that with strong effect sizes, following delays of at least 7 days, on an eight-trial *Long-Term Face Memory Test* (LTFMT8) SFRs ($n = 112$) as a group made significantly more target-present line-up hits, and target-absent CRs than average-ability controls ($n = 222$). Controls made more foil IDs. SFRs' sensitivity ($d^/$) and confidence was also higher. However, delay effects were weak and not significant, probably because most participants (87.1%) completed the LTFMT8 within a three-day period (7-10-days). Nevertheless, delay was not a confounder, as control and SFR delay did not differ. Figure 1 nonetheless shows large overlaps between SFR and control LTFMT8 scores, with some SFRs scoring below controls. It should be acknowledged that any 8-trial test will possess limited power to discriminate between high and low performers, and that the within-condition counterbalanced procedures differed. Data screening found this had no between-group impact, but differing task difficulty may have influenced individual scores. Experiment 3 employed same-condition procedures throughout to address this. However, it is very clear that a substantial proportion of short-term SFRs do not sustain their skills over longer retention intervals.

Line-ups comprised 6-person hybrid-videos or photos, and prior to viewing, half the participants were warned that the target, viewed in Phase 1 for 30s or 60s, ‘may or may not be present’. As expected, this warning induced caution (Clark, 2005). Hit rates were reduced; misses and CRs increased; but only in photo and not hybrid-video line-ups. Indeed, effects were dependent on face recognition ability, line-up media, and Phase 1 display time. Longer Phase 1 displays (60s *vs*. 30s) reduced misses, while hits were higher to hybrid-video than photo line-ups, particularly with longer Phase 1 display times. With longer exposure times, hybrid-videos may facilitate idiosyncratic movement cue recognition, or generate greater opportunity for recognition from multiple video frames. Nevertheless, these effects were mediated by a four-way interaction. For SFRs, the warning of potential target absence reduced hits to photo line-ups but not hybrid video line-ups. SFRs may therefore be disproportionally advantaged by movement in images, potentially indicative of more effective familiarisation during initial exposure to faces, and this provided immunity from the biasing effects of the warning. For controls, a significant three-way interaction was probed by inconclusive and possibly spurious non-significant post hoc analyses, suggesting that controls were more susceptible to the warning regardless of line-up media and target-presence.

The important implications for eyewitness researchers is that when conditions are optimised (i.e. Phase 1 of 60s *vs.* 30s; and hybrid-videos *vs.* photos) and participants are better face recognisers (i.e. SFRs *vs.* controls), reducing conditions of uncertainty, the warning has reduced impact. However, in sub-optimal eye-witnessing conditions, the warning has greater influence, and although the evidence here supports its use in reducing target-absent misidentifications, there is an increased risk of failing to identify a target. These results may also be the first to demonstrate that differences in face recognition ability may influence the impact of these standard line-up instructions.

The results have police line-up policy implications. The literature is divided as to whether simultaneous or sequential line-ups are more reliable (e.g. Mickes et al., 2012; Steblay, Dysart, & Wells, 2011; Wells et al., 2015; Wixted & Mickes, 2014). US police mainly employ simultaneous or sequential photo line-ups; UK police sequential video line-ups. The simultaneously displayed, but sequentially presented hybrid-video design combines elements from both. Future research should compare hybrid-video effectiveness with current UK and US systems. Large databases of volunteer foil videos have been captured for UK systems; and if the hybrid-video design proves advantageous, it should not be beyond software programmers' abilities to adapt formats.

Experiment 2

Experiment 1 revealed that as a group, SFRs exceeded controls at short- and long-term face recognition. Experiment 2 assessed repeated identification decisions, as many of Experiment 1's participants completed the line-up tasks a second time. Repeated identifications are associated with *commitment effects* (e.g., Hinz & Pezdek, 2001; Valentine, Davis, Memon, & Roberts, 2012; for a review see Deffenbacher et al., 2006). Regardless of accuracy, first procedure decisions, tend to be replicated. For instance, Valentine et al. (Experiment 3) found that most participants who viewed a live event and 15 minutes later identified either the target or an ‘innocent suspect’ from a live show up (one-person line-up); 1-30 days later *committed* to the same correct (95%) or incorrect (85%) identification from 9-person line-ups. Delay had minimal impact.

Negative commitments to making a ‘not present’ decision also occur (Deffenbacher et al., 2006), and foils viewed but *not* selected in a first procedure are more often than at chance rates likely to be selected in a second (Deffenbacher et al., 2006; Earles, Kesten, Curtayne, & Perle, 2008; Memon et al., 2003). These errors which tend to be rarer than commitment effects (see Blunt & McAllister, 2009; Goodsell et al., 2009), imply source confusion in associating the context in which the foil was first seen (e.g. Johnson, Hashtroudi, & Lindsay, 1993). Experiment 2 investigated if SFRs who might better remember such incidental encounters but not necessarily the context would make more such errors.

In addition, whereas ‘not present’ responses were correct in Experiment 1’s target-absent line-ups; in Experiment 2, participants were correctly informed all line-ups were target-present. Therefore, analyses were conducted on the subset of correctly rejected trials in Experiment 1, as to whether in Experiment 2, SFRs would be more likely to a) demonstrate stronger negative commitment effects by rejecting the line-ups again; b) select foils seen but not selected in Experiment 1, indicative of source confusion; or c) shift from correct Experiment 1 line-up rejection to correct Experiment 2 hits, suggesting immunity from commitment and source confusion effects. Overall, compared to controls, SFRs were expected to make more hits, and to display stronger commitment than source confusion error effects, by selecting more targets and/or foils previously selected in Experiment 1.

Method

Design

A mixed design compared SFR and control LTFMT8 decision consistency. The repeated measures factors were *Experiment 1 target-presence* (target-present *vs.* target-absent), and *line-up decision* (target *vs.* foil *vs.* not present). The primary dependent variables were Experiment 2 hit rates as well as *positive* (target, foil) and *negative* (not present) commitment effects as measured by rates of identical (*vs.* different) selections in both experiments.

Table 2: Age, gender, and ethnicity, and mean Experiment 3 LTFMT8 performances by all repeater participants, and comparisons between-group means (some demographic data were missing)

	All		SFRs		Controls					
n	539		57		103		*df*	*t*	*d*	*p*
Age	18-68 yrs.		18-49 yrs.		19-66 yrs.					
	M	*SD*	*M*	*SD*	*M*	*SD*				
	35.0	10.6	33.4	7.9	35.5	10.5	158	1.32	0.23	.189
Gender										
Male	197 (36.5%)		26 (45.6%)		42 (41.2%)			[A]		
Ethnicity										
White	86.2%		84.2%		95.1%			[B]		
	M	*SD*	*M*	*SD*	*M*	*SD*				
Experiment 3 LTFMT										
Delay	170.66	40.28	176.99	37.09	168.11	39.49	158	1.39	0.23	.166
Experiment 3 outcomes for Experiment 1 target-present trials										
Hits	0.34	0.28	0.50	0.28	0.27	0.22	158	5.54	0.91	<.001
Foil IDs	0.52	0.30	0.36	0.31	0.60	0.28	158	-4.96	0.81	<.001
Misses	0.13	0.20	0.14	0.18	0.13	0.22	158	0.01	0.05	>.2
Experiment 3 outcomes for Experiment 1 target-absent trials										
Hits	0.14	0.19	0.21	0.22	0.14	0.18	158	1.94	0.35	.054
Foil IDs	0.69	0.27	0.61	0.25	0.67	0.27	158	-1.29	0.23	.199
Misses	0.17	0.23	0.18	0.20	0.19	0.24	158	-0.20	0.05	>.2
Combined target-present and target–absent trials										
Confidence	43.89	19.11	48.67	19.14	42.52	18.47	158	1.99	0.33	.048

[A] Gender x group, χ^2 (1, 159) <1.

[B] Ethnicity (white = 1 *vs.* other = 0) x group, χ^2 (1, 333) = 5.39, *p* = .020, Cramer's V =.184.

Participants

Experiment 1 participants (*n* = 1,091) were invited to contribute to Experiment 2; 539 'repeaters' responded (49.4%). SFR (*n* = 57: 50.9%) and control (*n* = 103: 46.4%) repeater (*vs.* non-repeater) proportions were approximately equal, χ^2 < 1 (see Table 2 for demographic data). Independent-measures t-tests comparing repeater and non-repeater performances on Experiment 1's test outcomes found no significant CFMT+ or STFMT differences (all *t's*(1088) ≤ 1.79, *p* ≥ .074, *d* ≤ 0.11). However, repeaters had been more accurate (hits, CRs,

$d^{/}$) (all *t's*(1088) = 2.93, *p* ≤ .003, *d* ≥ 0.15), and confident, *t*(1088) = 2.26, *p* = .024, *d* = .014, on Experiment 1's LTFMT8.

Materials and procedure

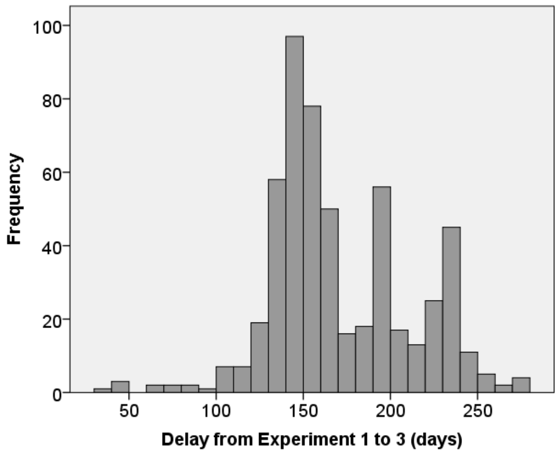

Figure 2: Delay for all participants between Experiment 1 and Experiment 2

The LTFMT8 Phase 2 procedure matched Experiment 1, except instructions stated that "the target-actor is definitely present in each line-up", and all 8 hybrid-video line-ups were target-present, although target and foil images were displayed in different positions to Experiment 1. Experiment 1 recruitment took a few months, with participation rate 'peaks and troughs', whereas the Experiment 2 link was open for about one month, resulting in substantial variations in delay between experiments (see Figure 1) (36-275 days, *M* = 170.7, *SD* = 40.3, skew = 0.24, SEM = 0.11, Shapiro-Wilk (539) = 0.949, *p* < .001).

Results

The full participant sample's results (*n* = 539), together with analyses of delay can be found in the General Discussion. Overall accuracy was poor. No participant achieved a maximum LTFMT score (*n* = 7 scored 7 out of 8; *n* = 73 scored 0; Median = 2, *M* = 1.94, *SD* = 1.38). Analyses examining the effect of Experiment 1 line-up media and warning conditions on Experiment 2 hits (Experiment 1 warning: Experiment 2: *M* = 0.24 *vs*. Experiment 1 no warning: Experiment 2: *M* = 0.24; Experiment 1 hybrid-videos: Experiment 2 *M* = 0.24 *vs*. Experiment 1 photos: Experiment 2: *M* = 0.24) found no reliable interactions with between-group effects (p > .2) and these conditions were collapsed. Separated by whether Experiment 1 trials were target-present or -absent, but regardless of Experiment 1 condition, Table 2 depicts Experiment 2 LTFMT8 mean outcomes by all participants, SFRs and controls, with independent-measures t-tests comparing groups.

Although not significant, mean SFR delay was 8 days longer than controls. SFRs made more Experiment 2 hits and fewer foil IDs than controls from line-ups previously target-present in Experiment 1. There were no miss rate differences, although there was also a marginally significant trend for SFRs to make slightly more hits to line-ups previously target-absent in Experiment 1.

Commitment effects: Regardless of target-presence, for the sub-set of LTFMT8 trials in which SFRs (n = 57) and controls (n = 100) made a selection from line-ups (target, foil) in Experiment 1 and 2, a 2 (group) x 2 (decision commitment: identical, different) x 2 (Experiment 2 accuracy: correct target, incorrect foil) mixed ANOVA compared the quantity of *identical* selections, indicative of commitment effects, and *different* line-up member selections; and whether Experiment 2 selections were of correct targets or foils (see Table S3 and Figure 3). Any Experiment 1 or 2 rejected line-up trials were excluded.

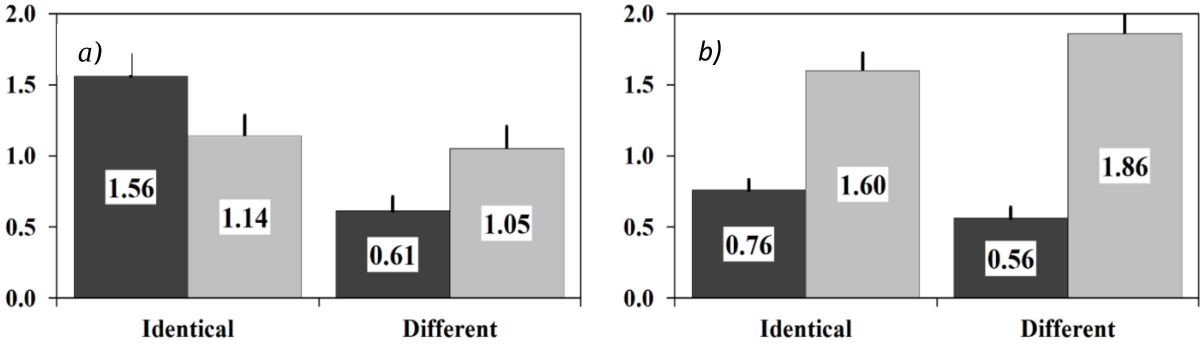

Figure 3: Mean Experiment 2 accuracy to identical and different line-up members for a) SFRs (n = 57), and b) controls (n = 100), from Experiment 1 to Experiment 2 (black bars = correct hits; grey bars = incorrect foil IDs in Experiment 2).

There were no group main effects. A significant decision commitment main effect, revealed more identical (M = 1.24, SD = 0.72) than different (M = 1.05, SD = 0.74) selections, supporting commitment effect theories and suggesting that even after approximately 6 months, participants remembered Experiment 1’s selected line-up member. A significant Experiment 2 decision main effect was due to more Experiment 2 selections being of foils (*M* = 1.53, *SD* = 0.83) than targets (*M* = 0.76, *SD* = 0.60), consistent with Experiment 1’s results and indicative of task difficulty.

The group x Experiment 2 accuracy interaction was significant. SFRs (*M* = 1.09, *SD* = 0.61) selected more targets than controls (*M* = 0.66, *SD* = 0.49). Controls (*M* = 1.73, *SD* = 1.10) selected more foils than SFRs (*M* = 1.10, *SD* = 0.70).

The critical group x decision commitment interaction was significant. SFRs (*M* = 1.35, *SD* = 0.67) and controls (*M* = 1.18, *SD* = 0.69) made similar numbers of identical selections. Controls (*M* = 1.21, *SD* = 0.77) made more different selections than SFRs (*M* = 0.83, *SD* = 0.70), suggesting reduced commitment effects.

The decision consistency x Experiment 2 accuracy interaction was significant. More identical selections were of exactly the same foils (*M* = 1.43, *SD* = 1.19) than of targets (*M* = 1.05, *SD* = 0.97). With larger effect sizes, more different selections were also of foils (e.g. change from correct ID to foil ID; or from one foil ID to a different foil ID) (*M* = 1.57, *SD* = 1.34), than changes from Experiment 1 foils to correct targets in Experiment 2 (*M* = 0.58, *SD* = 0.77), which may indicate a bias to guess when correctly informed the target was present.

The three-way interaction was not significant.

In summary, in the sub-set of line-ups in which participants made a selection in both experiments, SFRs were more likely than controls to consistently select the correct target. SFRs were also more likely to select an identical foil in both procedures, implicating stronger effects of commitment to identification decisions.

Negative commitment effects and source confusion: The final analyses examined numbers of Experiment 2 selections for the subset of correctly rejected Experiment 1 target-absent trials. Only participants who had made at least 1 correct rejection (out of 4 trials) in

Experiment 1 were included in a 2 (group) x 3 (Experiment 2 outcome: hits, foil IDs, misses) mixed ANOVA (see Table S4). The main effect of group was significant, $F(1, 121) = 6.35$, $p = .013$, $\eta^2 = .050$; purely a consequence of SFRs having made more correct target-absent decisions in Experiment 1. The Greenhouse-Geisser adjusted main effect of decision was significant, $F(1.68, 202.88) = 27.69$, $p < .001$, $\eta^2 = .186$. Indicative of source confusion, Bonferroni-corrected t-tests revealed foil ID rates were higher than line-up rejections indicative of negative commitment effects, which were in turn higher than hit rates indicative of immunity from both. The critical interaction was not significant, $F < 1$, suggesting no between-group negative commitment effects or source confusion susceptibility differences.

Discussion

Experiment 2's target-present hybrid-video line-ups were completed 36 to 239 days ($M = 170$) after Experiment 1's mixed target-present and -absent line-ups. Compared to controls ($n = 103$), SFRs ($n = 57$) made significantly more Experiment 2 hits and fewer foil IDs, to repeated target-present line-ups. They also made slightly more hits to targets originally target-absent in Experiment 1, and although this comparison was only marginally significant ($p = .054$), it does suggest superior ability to remember faces not seen for approximately six months.

Many participants identified the same target or foil selection in both experiments; with SFRs demonstrating stronger commitment effects of this type. As SFRs had made more Experiment 1 hits, commitment effects partly explain SFRs' higher Experiment 2 hit rates, although effects were not due to remembering selected faces' array positions, as these differed between experiments. As such, even after very substantial delays, SFRs more accurately remember faces previously selected than controls, although it is not possible to rule out factors such as clothing having an influence. This was identical in both experiments' line-ups, but was different from in the original Experiment 1 Phase 1 videos.

Analyses on Experiment 2 target-present line-ups previously correctly rejected when target-absent in Experiment 1 revealed no between-group negative commitment or source confusion differences. Suggestive of the influence of source confusion, for both SFRs and controls the only significant effects were that significantly more Experiment 2 foil identifications (56.9%) were made than not-present decisions indicative of negative commitments (28.3%), and target identifications (14.8%) respectively. However, as all line-ups contained five foils and one target, overall rates of correct target selections (14.8%) did not differ substantially from chance based on calculation of mean individual foil rates (56.9%/5 = 11.4%), meaning guessing cannot be ruled out. As such, supporting previous repeated line-up research (e.g. Blunt & McAllister, 2009), the evidence for source confusion is less compelling than positive commitment effects, particularly commitment effects in participants with superior face recognition ability.

Overall however, perhaps not surprisingly given that the designs of Experiment 1 and 2 were analogous to conducting eight eyewitness identification studies in one session, the LTFMT8 was susceptible to floor effects, and due to the counterbalanced conditions, individual analyses comparing each SR’s LTFMT performance with controls were not feasible. Experiment 3 addressed these issues.

Experiment 3

Experiment 3 employed a 10-trial version of the LTFMT with mean delays of between 1 and 63 days. To reduce floor effects, all Phase 1 videos were 60s; only the hybrid-video line-ups were employed, and participants were correctly informed all line-ups were target-present. Participants completed face memory ability self-assessments, the CFMT+ (Russell et al., 2009), the simultaneous *Glasgow Face Matching Test* (GFMT) (Burton et al., 2010), and Experiment 1’s STFMT immediately prior to the LTFMT10 Phase 1.

The inclusion of the GFMT, a perceptual task with no memory demands, allowed replication of previous research finding that some SFRs possess poor simultaneous face matching skills, while others are outstanding at both face memory and perceptual tests (e.g. Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Bobak, Dowsett et al., 2016; Bobak, Hancock et al., 2016; Davis et al., 2016, see Davis & Valentine, 2015 for a review of simultaneous face matching). Effects may be due to task demands, in that accurate face matching relies on feature-by-feature strategies (e.g. Megreya & Burton, 2006); whereas face memory draws more on a holistic whole-face approach (e.g. Tanaka & Farah, 1993). Indeed, similar dissociations are found in some developmental prosopagnosics and brain-damaged acquired prosopagnosics (Bate, Haslam, Jansari & Hodgson, 2009; De Haan, Young, & Newcombe, 1987, 1991).

In addition to between-group analyses, individual analyses compared performances of each SFR against control means, allowing between-test consistency measurement and to generate an estimate of the proportion of the general population each SFR would be expected to exceed. The hypotheses mainly replicated previous experiments. SFRs were again expected to outperform controls, with longer delays predicted to reduce LTFMT10 accuracy.

Method

Design

In Experiment 3, after viewing 10 1-min target-actor videos in Phase 1, participants randomly received invites to Phase 2 of a 10-trial target-present hybrid-video version of the LTFMT10 after delays of 1, 7, 14, 28 or 56 days. Between-group and individual analyses (Crawford, Garthwaite, & Porter, 2010) compared SFRs and controls.

Participants

Participants contributed after reading media articles linked to an online anonymised 5-min, "*Could you be a super-recogniser?" Test*. On debriefing, invites described the current study. In total, *n* = 9,715 participants clicked on the study link and provided consent; although many dropped out (*n* = 9,118), mostly after the GFMT (CFMT+ completers: *n* = 4,403, STFMT: *n* = 3,250, GFMT: *n* = 3,070, LTFMT10: *n* = 597). Note that only participants completing the GFMT are included in any reported analyses (*n* = 3,070; male = 1,119 (36.4%), female = 1,951; white-Caucasian = 2,391 (81.0%), aged 16-74 years, *M* = 32.5, *SD* = 11.0).[3] None were compensated or participated in other experiments.

Two one-sample t-tests revealed that included participants (*n* = 3,070) outperformed Bobak, Pampoulov, et al.'s (2016) CFMT+ (*M* = 83.23, *SD* = 11.12), *t*(3069) = 62.44, *p* < .001, *Cohen's d* = 1.47, and Burton et al.'s (2010) GFMT (*M* = 36.63, *SD* = 2.77), *t*(3069) = 82.63, *p* < .001, *Cohen's d* = 1.48 and norms respectively.[4]

Initial SFR (95/102) and control CFMT+ thresholds (58-83/102) replicated Experiment 1. The STFMT ($d^{/}$) verified ability (SFR: *n* = 414, *M* = 2.20, *SD* = 0.73 *vs.* control: *n* = 1,204, *M* = 1.43, *SD* = 0.54), although the mid-point between $d^{/}$ means used as the final maximum and minimum thresholds respectively was slightly lower than in Experiment 1 ($d^{/}$ = 1.815 *vs.* 2.085) (see Table S5 for final group allocation and demographic data). Lower mean STFMT scores in Experiment 3 than Experiment 1 may be a consequence of fatigue from prior CFMT+ completion, as well as that Experiment 1's participants had received a specific e-mail invite. Therefore, Experiment 3's slightly different inclusion criteria were retained.[5]

Materials

[3] Some demographic data are missing. The next largest self-defining ethnic group were from the Asian sub-continent/British-Asian (*n* = 82, 2.8%).

[4] Two one-sample t-tests revealed that LTFTM10 completers (*n* = 597) outperformed Bobak, Pampoulov, et al.'s (2016), CFMT+ (*M* = 86.62, SD = 9.52) *t*(596) = 40.87, *p* < .001, *Cohen's d* = 1.45, and Burton et al.'s (2010), and GFMT (*M* = 37.32, *SD* = 2.46) *t*(596) = 47.83, *p* < .001, *Cohen's d* = 1.48 norms respectively. Effect sizes were larger than with the whole sample.

[5] If Experiment 1's criteria had been exactly replicated, the Experiment 3's SFR group would have been reduced from *n* = 84 to *n* = 76; the control group from *n* = 103 to *n* = 98. However, between-group analyses conducted with and without these participants revealed no impact on main effects, interactions or post hoc tests. The individual results of super-face-recognisers not meeting Experiment 1's criteria on individual analyses are marked in Figure 5. Again however, there was no obvious pattern that differed from the remaining SFRs.

Glasgow Face Matching Test (short version) (GFMT) (Burton et al., 2010): This 40-trial test contains simultaneously displayed pairs of white-Caucasian male and female facial images. Twenty trials are matched (correct response: “same” - hit); 20 mismatched (“different” - CR). Burton et al.’s participants’ (*n* = 194) performances ranged from 20-40 out of 40 (*M* = 32.5 (81.3%, *SD* = 9.7)). Current participants’ scores (*n* = 3,070) ranged from 21-40 (*M* = 36.6 (91.6%, *SD* = 6.9).

Long-Term Unfamiliar Face Memory Test (10-trial: LTFMT10): This version replicated Experiment 1, except additional Phase 1 and 2 male and female target-actor videos and target-present hybrid-video line-ups (Actors I and J) were displayed immediately after the original eight. All 10 Phase 1 videos were displayed for 60s, and participants were correctly told that, “the target-actor is definitely present in each line-up”.

Procedure

After providing informed consent, and demographic data, participants completed face recognition ability self-assessments (1-5), followed by the CFMT+ (Russell et al., 2009), STFMT, and GFMT (Burton et al., 2010). Participants were then fully debriefed and provided with their scores on the tests above, and at the same time asked to contribute to the LTFMT10, and to provide e-mail addresses in order to be sent the link to Part 2. If they provided consent, they viewed LTFMT10 Phase 1. Providing e-mails removed anonymity which may be why there was a large drop out (only *n* = 1570 out of 3070 (51.1%) reading the invite participated). On completion they were warned to watch out for their Phase 2 invite e-mail, although as delay was random no information was provided as to when it would be sent.

In total, 1570 participants completing LTFMT10 Phase 1 were e-mailed Phase 2 invites. A significant chi-squared test, $\chi^2(4, 1570) = 15.77$, *p* = .003, Cramer’s V = 0.100, revealed that the highest proportion took part (1 = participated, 0 = did not) following the one-day invite (1-day: 45.3% responded; 7-days; 39.7%; 14-days: 32.4%; 28-days: 32.9%;

56-days: 39.7%). A one-way ANOVA examining participants' response delay between each delay condition was significant, $F(4, 592) = 8.08$, $p < .001$, $\eta^2 = .052$. The longest mean delays after the e-mail was sent before participation were in the 56-day condition (1-day: 1.94 days, 7-days: 2.83 days, 14-days: 3.62 days, 28-days: 2.90 days, 56-days: 7.94 days).

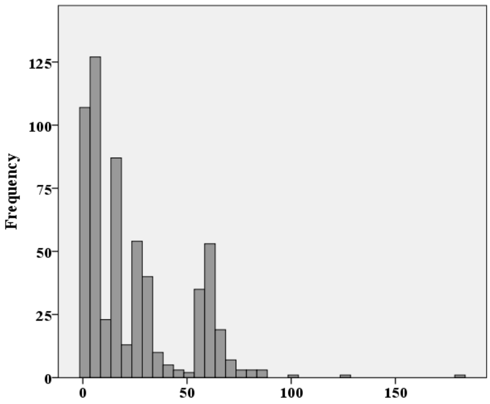

Figure 3: Mean delay for all participants for Experiment 3 (n = 597)

A chi-squared test, $\chi^2(1, 534) = 19.28$, $p < .001$, Cramer's V = 0.190 found more SFRs (48.0%) completed Phase 2 than controls (28.7%). As with Experiment 1, some participants (n = 203; 12.9%) reported problems with Phase 1 videos, so that 372 out of 15,700 (2.36%) were replayed. This had no impact on whether participants completed Part 2 or not; or on results reported below ($p > .2$).

Results

Data of all participants can be found in the General Discussion. Here only SFR and control data are reported if they completed the LTFMT10. LTFMT10 delay varied from 1-182 days (see Figure 3); and was skewed (Shapiro-Wilk (597) = 0.828, $p < .001$), although it did not differ between SFRs and controls, $t(185) < 1$. Seven participants achieved LTFMT10 scores of 10/10 (n = 21 scored 0).

Table S5 displays mean group performances on all tests, including, regardless of delay, the LTFMT10. Independent-measures t-tests demonstrate that with strong effect sizes, SFRs outperformed controls on all outcomes except decision confidence, while displaying a liberal response bias to respond ‘old’ on the STFMT.

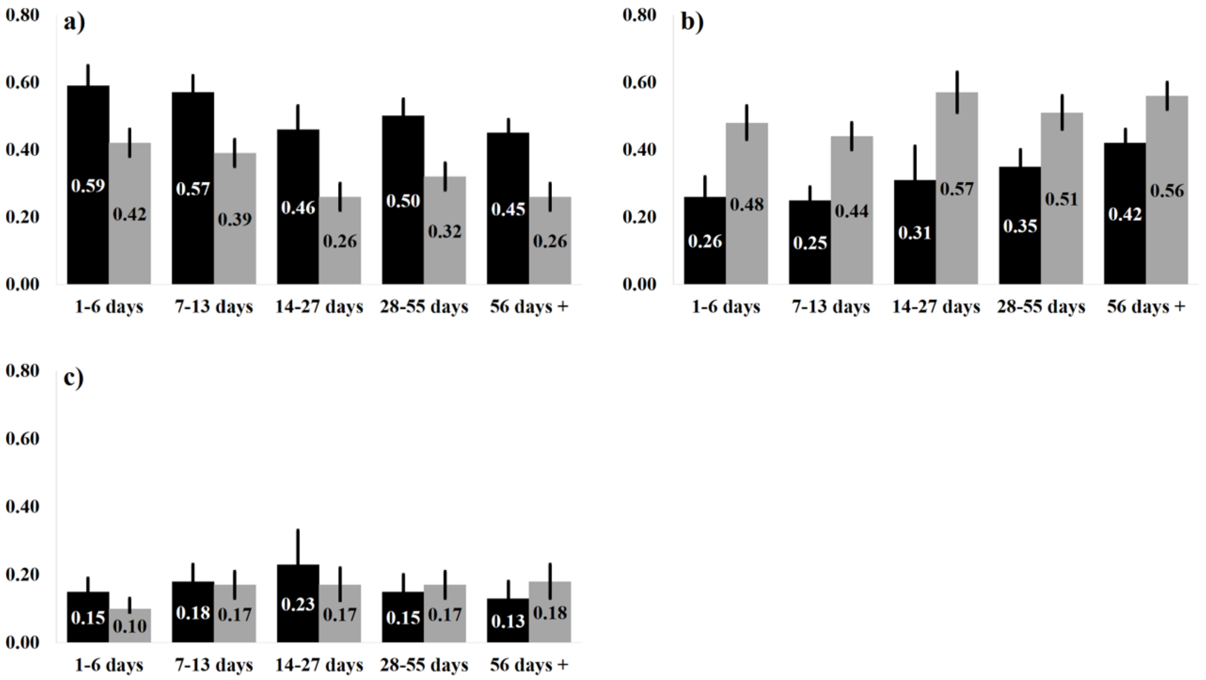

Figure 4: LTFMT10 (a) Hit, (b) foil ID and (c) miss rates in Experiment 2 as a function of group and delay (dark bars = SFRs, grey bars = controls)

Analyses examined actual delay (regardless of randomised invite group) and group influence on the LTFMT10, in three 2 (group: SFR, control) x 4 (delay condition; 1-6 days (M = 1.52), 7-13 days (M = 7.75), 14-27 days (M = 15.82), 28-55 days (M = 31.01), 56+ days (M = 64.21)) ANOVAs on hits, foil IDs and misses (see Figure 2). Significant group main effects were found with LTFMT10 hits, $F_{Hits}(1, 177) = 36.63$, $p < .001$, $\eta^2 = .171$, and foil IDs, $F_{Foil\ IDs}(1, 177) = 36.63$, $p < .001$, $\eta^2 = .170$; but not misses, $F_{Misses}(1, 177) < 1$. SFRs made more hits and fewer foil IDs than controls.

Significant delay main effects were found with hits, $F_{Hits}(4, 177) = 4.34$, $p = .002$, $\eta^2 = .089$, and foil IDs, $F_{Foil\ IDs}(4, 177) = 3.38$, $p = .011$, $\eta^2 = .071$, but not misses, $F_{Misses}(4, 177) < 1$. Tukey’s tests found hit rates were higher after 1-6 days, and 7-13 days compared to 56+ days ($p < .05$ both comparisons). Foil IDs were higher with 56+ delays than 7-13 days ($p < .05$). No other comparisons or interactions (all $F’s(4, 177) < 1$) were significant ($p > .05$).

Confidence: The final analyses examined mean confidence in hits (n = 576, M = 62.9, SD = 23.2), foil IDs (n = 570, M = 63.2, SD = 33.3), and misses (n = 308, M = 52.7, SD = 27.5). A 2 (accuracy: correct (hits), incorrect (foils, misses)) x 2 (group) x 5 (delay) mixed ANOVA revealed only a significant group effect, $F(1, 167) = 6.95$, $p = .009$, $\eta^2 = .040$. [6] As reported in Table S5, SFRs (n = 79) reported higher confidence than controls (n = 98).

Individual level analyses: Modified t-tests for single cases (Crawford et al., 2010), individually compared the scores of SFRs against the control mean on the CFMT+ (out of 102), the STFMT ($d^/$), the GFMT ($d^/$) and the LTFMT10 (hits). Figure 5 a-d depicts 95% confidence intervals of the proportion of the general population each SFR would be expected to exceed. Even though groups were created based on these tests, CFMT+ and STFMT figures are displayed as they serve to demonstrate outcome differences, with large effect sizes compared with the GFMT and LTFMT10.

Not surprisingly given inclusion criteria, all SFRs (n = 84, 100%) scored significantly higher than the control mean on the CFMT+, $t(103) = 3.12\text{-}4.23$, $p < .05$, one-tailed, $z = 3.14$ (95% CI: 2.67-3.61) - 4.25 (95% CI: 3.64-4.87); and the STFMT ($d^/$), $t(103) = 1.84\text{-}10.32$, $p < .05$, one-tailed, $z = 1.84$ (95% CI: 1.52-2.15)-10.37 (95% CI: 8.93-11.80). Most exceeded the control GFMT ($d^/$) mean (n = 77, 91.7%), 32 significantly (38.1%), $t(100) = 1.89$, $p < .05$, one-tailed, $z = 1.90$ (95% CI: 1.57-2.20) (i.e. with scores higher than 95% of the estimated population), although this required a score of 100%. Noteworthily, 7 SFRs (8.3%) scored below the control GFMT mean.

For the LTFMT10, SFRs were compared to controls from the same delay groups. The raw scores of 69 SFRs (82.1%) exceeded the associated delay-condition control mean, although only 17 (20.2%) significantly (a further 5 were only one score below), $t(84)$ = 1.90-

[6] Note: A similar mixed ANOVA with confidence in hits, foil IDs and misses separated so that participant numbers were lower as fewer provided a response on each type (n = 85), only revealed a significant response type effect, $F(2, 150) = 11.63$, $p = .009$, $\eta^2 = .134$. Confidence in hits and foil IDs did not differ ($p > .05$), but both were significantly higher than confidence in misses ($p < .05$).

4.03, $p < .05$, one-tailed, $z = 1.94$ (95% CI: 1.29-2.58)-3.29 (95% CI: 2.60-5.70), while 15 SFRs scored below control means. From Figure 5 it can be seen that the worst SFR performers on the LTFMT10 were not necessarily the same participants as poor scorers on the GFMT, providing more evidence for perceptual and memorial skill dissociation.

Figure 5a: CFMT+ (Controls (n = 103) M = 75.33, SD = 6.27)

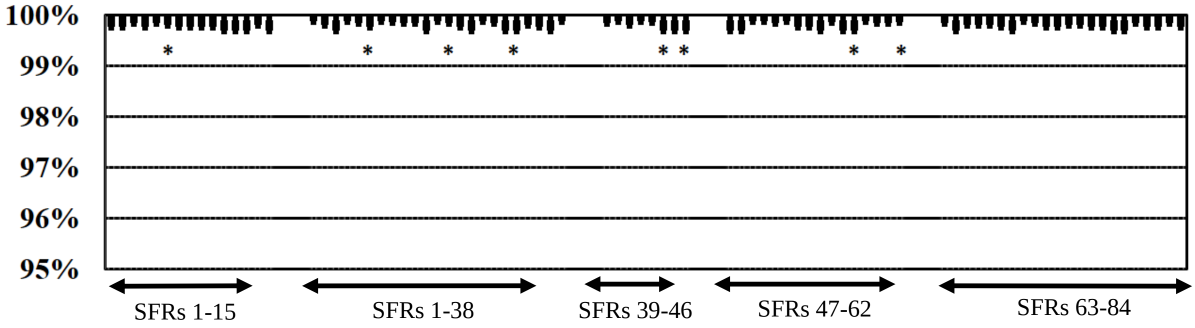

Figure 5b: STFMT (d/) (Controls (n = 103) M = 1.38, SD = 0.25)

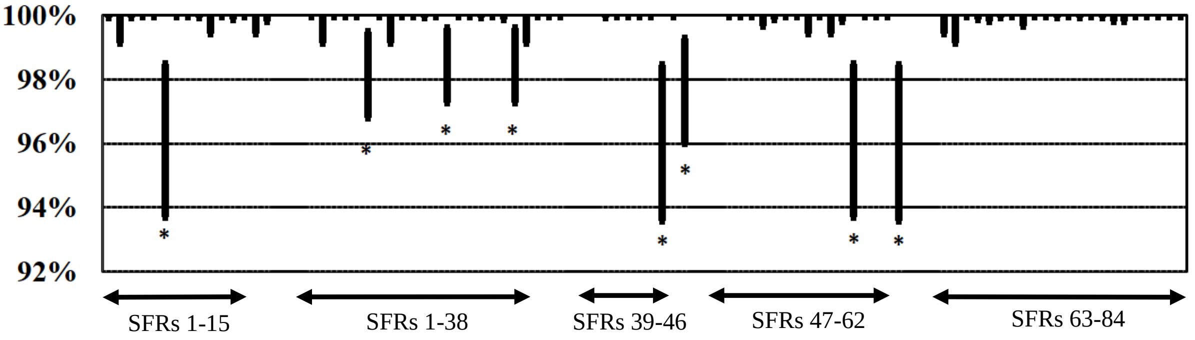

Figure 5c: GFMT (d/) (Controls (n = 103) M = 2.64, SD = 0.68)

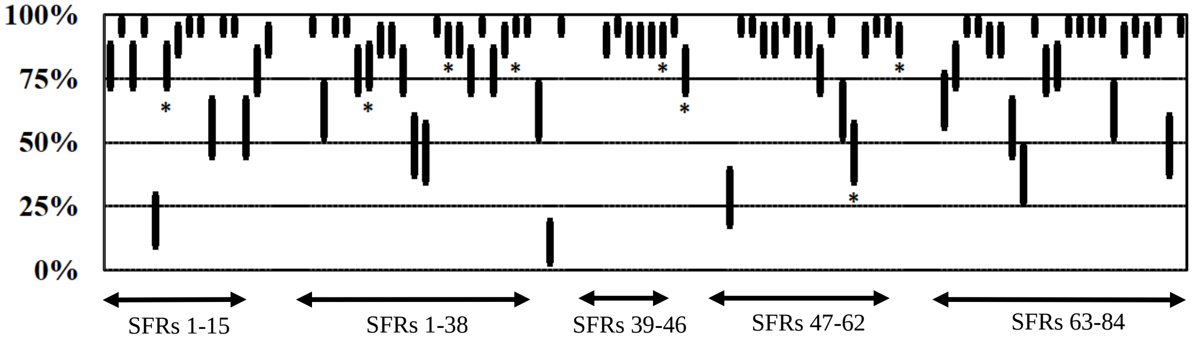

Figure 5d: LTFMT10 (Controls: 1-6 days (n = 19, M = 0.42, SD = 0.17); 7-13 days (n = 27, M = 0.39, SD = 0.21); 14-27 days (n = 16, M = 0.26, SD = 0.15); 28-55 days (n = 16, M = 0.32, SD = 0.16); 56+ days (n = 25, M = 0.26, SD = 0.20)

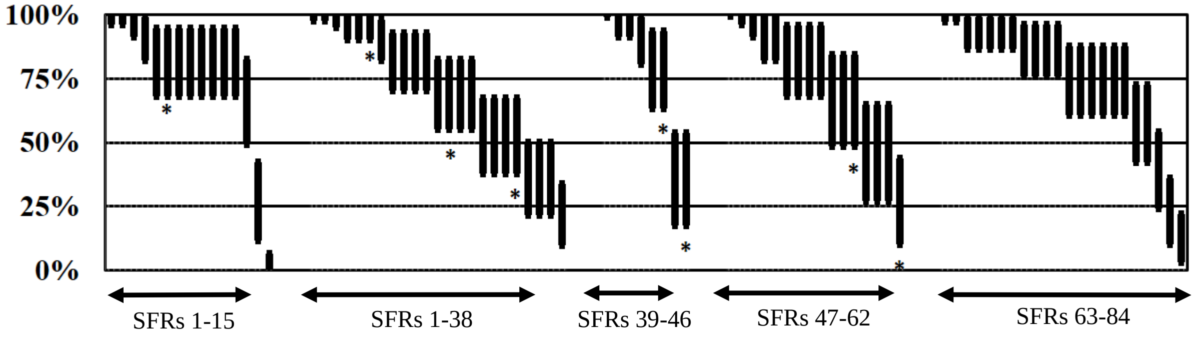

Figure 5: Upper and lower bound confidence intervals (95%) of the estimated proportion of the population expected to fall below each SFR (n = 84) based on (5a) CFMT+ scores, (5b) STFMT d/, (5c) GFMT d/, and (5d) LTFMT10 hits. Due to low SFR variability on the CFMT+ and STFMT, only the 95%-100% and 92%-100% range is depicted in Figures 5a and 5b. The 50% line on Figures 5c and 5d represents the control mean, so that 50% of the population would be expected to achieve above this level. To enhance interpretability throughout, SFRs are grouped based on delay condition (1-6 days, 7-13 days, 14-55 days, 56+ days), and rank-ordered from left-to-right based on LTFMT10 hit rates.
** Experiment 3 SFRs not achieving Experiment 1 SFR criteria on the STFMT (see footnote 5).*

Discussion

In Experiment 3, planned delays on the target-present 10-trial LTFMT10 varied between 1 day and 56 days, and as expected, longer delays were associated with decreased hit rates, and increased foil IDs although effect sizes were small and only a few post-hoc comparisons were significant. With far stronger effect sizes, as a group, SFRs were more accurate and confident than controls in all delay conditions. Importantly, there was no interaction between group and delay. As such, the shape of SFRs and controls' forgetting curve appears similar, and SFRs' superiority over controls after 56 days is of approximately the same degree as after 1 day. SFRs' face memory advantage appears driven by superior encoding, as well as immediate, and longer-term retention.

Individual analyses demonstrated that 100% (*n* = 84) of SFRs significantly exceeded the control mean on the CFMT+, and the STFMT (d/), not surprisingly given that scores on these tests were used for group allocation. Most SFRs also exceeded the control mean on the GFMT (*n* = 77, 91.7%); and the LTFMT (*n* = 69, 82.1%). However, only a minority significantly outperformed controls on these tests (GFMT: *n* = 32: 38.1%; LTFMT: *n* = 17: 20.2%) (see Figure 5). This may partly be due to low test discriminatory power. SFR GFMT mean (38.6 out of 40), was only slightly above controls (*M* = 35.5), and for individual analyses to be significant, a maximum score was required. Nevertheless, as found previously (e.g. Bobak et al., 2016), a few SFRs performed below the control mean on the GFMT (*n* = 7, 8.3%), supporting propositions for dissociated perceptual and memorial SFR components (e.g. Bobak, Dowsett et al., 2016). Of course, poor performance on a single test may be due to extraneous factors (e.g. distractions, internet disruptions, lack of motivation etc.).

This was also the first research to demonstrate SFR variability at longer term face memory as a substantial minority performed below the LTFMT10 control mean (*n* = 15, 17.8%). With 10 trials this test had even lower discriminatory power than the GFMT, so that regardless of delay SFR hit rates (*M* = 0.52) were only slightly higher than controls (*M* = 0.33). In addition, the confounding factors of distractions and internet disruptions were also more likely to have an adverse influence with the strictly timed Phase 1 video display times and far longer delays. Large numbers of participants of all abilities, but particularly controls additionally failed to complete the LTFMT10 particularly in the longer delay conditions (*n* = 1570 completed Phase 1; *n* = 597 completed Phase 2: 38.0%) (52.0% of SFRs and 71.3% of controls dropped out), and it is uncertain whether the pattern of results would have differed if more had finished. On the other hand, the results match Experiments 1 and 2, and the investment in time of those who did finish was high, and it would be surprising for someone believing they possessed superior ability based on short-term test scores which were provided at the end of the GFMT to take part but not try. As such, these results suggest that for inclusion in research, or police SFR groups, long-term face memory and simultaneous face matching tests should be included to ensure possession of a full superior skill battery.

General Discussion

The three experiments described here demonstrate that SFRs whose scores on two short-term face memory tests would probably have met any previous super-recogniser definition mainly possess above average long-term face memory as well. Delays varied from 1 day to approximately 6 months. On the *Long-Term Face Memory Test* (LTFMT), compared to average-ability controls, SFRs as a group made more target-present line-up hits, and target-absent line-up correct rejections (CRs). Between-group effect sizes were stronger than the effects of delay. Similar between-group simultaneous face matching effects were found in Experiment 3. These findings are consistent with models suggesting individuals in the upper end of the human face recognition ability spectrum possess higher-order, face-specific visual memory system mnemonic enhancements, and that these can also be sustained over long retention intervals. Nevertheless, as with previous research (e.g. Bobak, Dowsett et al., 2016), there was substantial SFR variability, with some performing below control means on the LTFMT and the Glasgow Face Matching Test (GFMT) (Burton et al., 2010).

In Experiment 1, with delays of at least a week, longer Phase 1 display times, and hybrid-video line-ups rather than photo line-ups facilitated higher accuracy and confidence in SFRs (*n* = 112) than controls (*n* = 222). Movement mainly advantages familiar face memory (e.g. Lander & Chuang, 2005), and these results suggest more efficient face familiarisation in SFRs, so that movement as an identity cue transfers to newly established face representations. SFRs were also less susceptible to the conservative response bias typically induced by warnings that line-ups ‘may or may not contain the target’, suggesting that the interpretation of these instructions designed to reduce misidentification risk in real police line-ups may partly be dependent on face recognition ability and confidence in identifications.

In Experiment 2, Experiment 1 participants (*n* = 539) identified targets a second time from target-present line-ups. Delay varied substantially (36-275 days, *M* = 170 days). With large effect sizes and higher confidence, SFRs (*n* = 57) outperformed controls (*n* = 103), particularly when Experiment 1 trials had also been target-present, although there was a marginal non-significant trend for higher SFR hit rates from previously target-absent line-ups as well (p = .054). SFRs also demonstrated stronger positive commitment effects to re-selecting the same targets *and* foils from Experiment 1. However, there was no reliable evidence of between-group negative commitment effects or source confusion.

With mean delays varying from 1 to over 56 days, only target-present line-ups were employed in Experiment 3. Between-group effect sizes were larger than those for delay, and although there were substantial individual differences within each delay condition, most SFRs (n = 66 out of 84, 82.1%), exceeded the control mean on the LTFMT10 (n = 100), 17 significantly (20.2%). These smaller proportions are perhaps not surprising given the low discriminatory power of a 10-trial test. Nevertheless, with no interaction, the shape of SFRs and controls forgetting curve was also similar (Deffenbacher et al., 2008), suggesting the driver of SFRs' superior skills is initial familiarisation, although they may also possess sustained larger face storage capacity (see Bobak, Dowsett et al., 2016; Jenkins, Dowsett, & Burton, 2018). Indeed, the mean super-face-recogniser hit rate after 56 days (M = 0.45) was roughly the same as the mean control hit rate (M = 0.42) after delays of between 1-6 days.

Predictors of long term face memory accuracy

To examine which factors best predicted long-term memory performance, the data from the three experiments were combined, with correlation coefficients conducted between each test (Table S6). Between-condition data were collapsed. Scores on the CFMT+, GFMT and STFMT were all significantly correlated with LTFMT accuracy outcomes in each experiment, although coefficients were often relatively small, a probable consequence of the recruitment bias to attract far better-than-average performers to this research reducing representativeness, and restricting score range. Indeed, in all experiments, participants' scores on the CFMT+ (n = 4458, M = 83.7/102 *vs.* 70.7) (Bobak, Pampoulov et al., 2016; and GFMT (n = 4451, M = 36.8/40 vs. 32.5)[7] (Burton et al., 2010) were more than 1 SD above

[7] GFMT data of 7 Experiment 1 participants were missing.

norms. Nonetheless, three multiple regressions were conducted with each experiment's LTFMT primary accuracy measure as the dependent variable (Experiment 1: $d^{/}$; Experiment 2 and 3: hits), with within-experiment conditions again collapsed. Predictors were pre-test face memory ability self-beliefs, GFMT (scores out of 40), CFMT+ (scores out of 102), and STFMT accuracy ($d^{/}$), as well as LTFMT delay and confidence; while Experiment 1's LTFMT was included as a predictor for Experiment 2's repeated line-ups.

Influential predictors differed by experiment (Tables S7-9),[8] although throughout STFMT ($d^{/}$) predicted LTFMT accuracy with stronger effects than the GFMT and the CFMT+. Indeed, the CFMT+ was only a significant predictor of LTFMT accuracy (hits) in Experiment 3. These effects are possibly due to the similarly-structured old-new designs of the SFTFMT and LTFMT, in that both require learning a series of hairstyle-included faces. In contrast, participants are familiarised to six hairstyle-removed target faces with over repeated trials in the CFMT+, which may generate a very different learning experience.

Face recognition ability self-beliefs did not predict long-term face memory accuracy, and only weakly correlated with accuracy on the other tests. This was perhaps not unexpected given the recruitment bias to attract better face recognisers, and that previous research in this area has been conflicting (e.g. Bate et al., 2018; Palermo et al., 2017). Nevertheless, of participants who met combined-Experiment 1 and 3 SFR criteria (see below) (*n* = 422), scores on the 5-point scale (well above – well below average) varied (see Table S10), with some providing surprisingly low appraisals of their actually good ability. It is clear that for job roles in which superior ability is essential, empirical performances on tests should be weighted far higher than self-belief in ability.

LTFMT accuracy was also predicted by confidence in Experiments 1 and 3, but not Experiment 2, partly supporting the confidence-accuracy relationship often found in

[8] Collinearity and multicollinearity assumptions were met (tolerance was 0.63 to 0.99, VIF 1.01 to 1.65).

eyewitness research (for a review see Sauer & Brewer, 2015). The non-significant Experiment 2 findings may be a consequence of the repeated line-up procedure. Not surprisingly, considering the strong effects of commitment found in Experiment 2, Experiment 1's LTFMT $d^{/}$ was the primary predictor of hit rates on Experiment 2's LTFMT. Delay only predicted LTFMT accuracy in Experiment 3, possibly due to the Phase 2 e-mails being sent after specific randomised delays of 1, 7, 14, 28 or 56 days enhancing variability, whereas in Experiment 1 and 2, despite the very substantial retention interval range, many participants completed Phase 2 after roughly similar delays. It is clear that from the weak delay effects found throughout, and in these regression analyses that individual differences in face recognition ability has a stronger impact on performance than retention interval.

Defining super-recognition ability

Most previous research has primarily employed the CFMT+ to assign superior-recognisers to groups. Here, a substantial proportion of current participants (≈ 30%) who would have been included in a SFR (*n* = 601) group based on Bobak, Pampoulov et al.'s (2016) rigorous CFMT+ criteria alone (≥ 95), were excluded following STFMT verification, although due to small between-experiment differences in STFMT performances, inclusion thresholds slightly differed. The data reported in Experiment 3 and in particular Figure 5 suggest that these between-experiment criteria had little impact.

Despite this, out of the 422 SFRs who met a combined Experiment 1 and 3 CFMT+ (≥ 95) *and* STFMT ($d^{/} \geq 1.88$) threshold calculated using the same formula as described previously, but based on all 4,458 participants, scored below control means on the GFMT (*n* = 20 out of 422, 4.7%) and/or LTFMT (combined Experiment 1 and 3 versions) (*n* = 65 out of 208, 31.3%). This is perhaps not surprising as these tests possess low discriminatory power with only 40 trials on the GFMT (which suffers from ceiling effects), and either 8 or 10 on the LTFMT (which suffered from floor effects). Although only one SFR scored below the established norms on the GFMT (Burton et al., 2010), these figures support previous research finding that some super-recognisers are relatively poor at this perceptual task, matching the dissociations found in developmental and acquired prosopagnosia (Bate et al., 2009; De Haan et al., 1987, 1991). Furthermore, and importantly for policing and security, it is clear that a far greater proportion of individuals who easily meet short-term criteria for super-recognition cannot sustain these abilities over longer retention intervals.

On the other hand, of the cross-experiment sample of 422 SFRs, 140 (33.3%) achieved 100% on the GFMT (another 122 (29.0%) scored 39/40); while a smaller proportion achieved scores that were significantly higher than controls ($p < .05$) on the different versions of the LTFMT. From an applied perspective, consistently exceptional face processing ability might benefit law enforcement. Indeed, some police forces have established super-recogniser units, and although it is not always clear what criteria selection was based on (e.g., Davis et al., 2016; 2018), any threshold will be somewhat arbitrary, and indeed, tests can only provide a marker of ability. They cannot guarantee actual workplace performance. Nevertheless, the results of the current research suggest that recruitment criteria could best be based on achieving test scores measuring different skills within a pre-tested super-recogniser group's typical range, rather than high performances on a single short-term memory test.

The absence of any gold-standard marker means that arguments can be made for different thresholds. A generic criteria commonly used to classify an individual as a likely member of an alternative (non-control) group is a score >2 SD above the normative control mean (Crawford & Garthwaite, 2012) on all tests. This threshold ensures that less than 2.5% of genuine controls would be classified as super-recognisers, and has been the standard commonly employed in super-recognition research (e.g. Bobak, Bennetts et al., 2016). It was also the initial CFMT+ criteria used here. However, due to ceiling effects on the GFMT

(Burton et al., 2010), this would exclude everyone, as even a maximum score of 40 out of 40 (z = 1.17, p > .05) would not achieve threshold (Control: M = 36.64, SD = 2.87).

Two-hundred and seven (out of 422) SFRs completed the LTFMT in Experiments 1 or 3. Figure 6 provides the normative z-scores of those meeting combined short-term experiment criteria described above. Although, the number of individuals meeting the following criterion rapidly diminishes with an increasing number of tests as correlations decrease (e.g. McDonald, 1999), an alternative but more stringent super-recogniser criterion might be to assign membership if an individual scores above the SFR normative mean score on *all* tests (i.e. z-scores > 0 in Figure 6). Only 18 out of 208 (8.7%) SFRs achieved this threshold (0.11% of 1688 all-ability LTFMT completers).

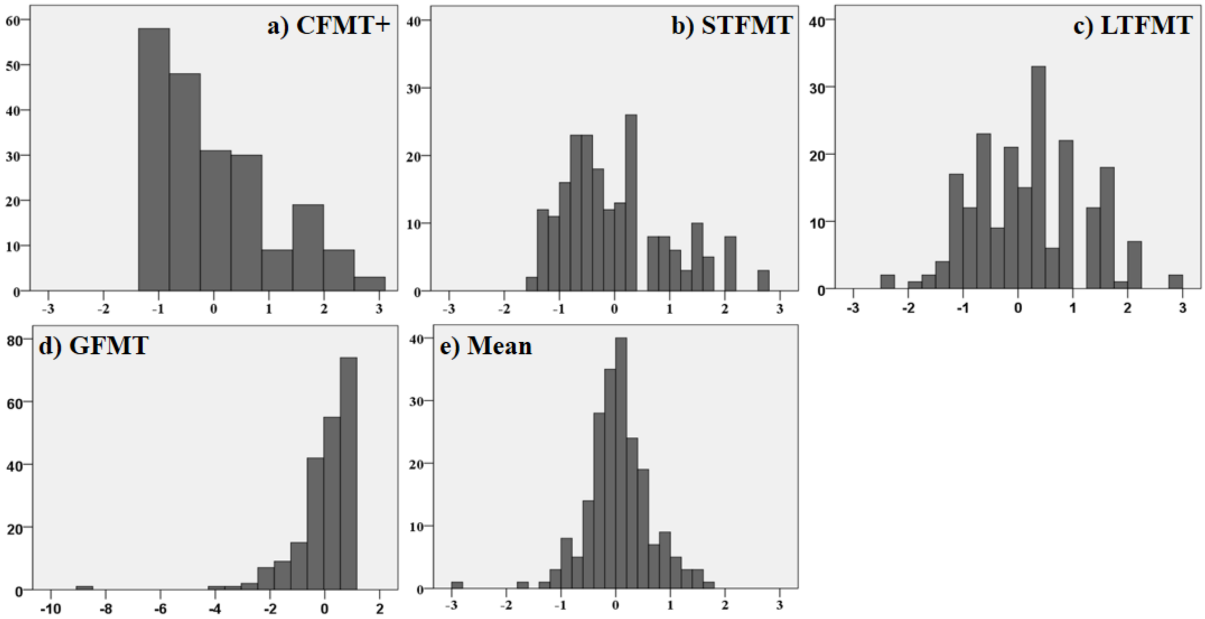

Figure 6: Frequency distributions of normative z-scores of SFRs (n = 208) who completed the LTFMT [(a) CFMT+: M = 96.94, SD = 1.79; (b) STFMT (d/): M = 2.61, SD = 0.49; (c) combined LTFMT: Experiment 1 (d/): M = -0.55, SD = 1.68, Experiment 3 (hits): M = 0.53, SD = 0.22; (d) GFMT: M = 38.56, SD = 1.66] as well as (e) mean z-scores calculated from a-d.

One compromise might be to select participants whose z-scores fall within one standard deviation of the SFR normative mean (i.e. z-scores on all tests > -1) which may help reduce rejection of potentially useful individuals, but still include those who score within the 'typical' super-recogniser range. One hundred and six achieved this criterion (0.62% of 1688). A final criteria would be to select participants whose *mean* z-scores across all tests were above zero. This was achieved by 130 (0.77% of 1688) (see Figure 6e). Nevertheless, all four criteria would result in the selection of far fewer super-recognisers than expected if 1-2% of the population possess this ability (see Russell et al., 2009).

Limitations and recommendations for future research

Noyes et al. (2017) recommend that individual analyses such as that conducted in Experiment 3 should form the primary assessment of super-recogniser abilities. However, it would not have been feasible to conduct this type of analyses in Experiments 1 and 2 due to the large numbers of counterbalanced conditions. Yet, this experiment adds to the knowledge of the nature of superior-face-recognition ability. We acknowledge that although the between-groups results may not be as robust as with individual analyses, they demonstrated that SFRs may use different decision-making processes when interpreting instructions and viewing line-ups, outcomes that would not have been otherwise apparent. Indeed, this research may have important implications in terms of policy for police eyewitness procedures. In Experiment 1, accuracy was higher to the novel hybrid-video design than photograph simultaneous line-ups, and future research could compare its efficacy to other video line-up types. It may prove superior as it combines the advantages associated with simultaneous displays (e.g. Mickes et al., 2012) and movement (e.g. Havard et al., 2010).

The LTFMT employed here also possessed low discriminatory power with a relatively small difference between mean scores of SFRs and controls. Despite this, trial numbers were far higher than most previous long retention interval face recognition or eyewitness identification research, and some participants produced very high scores. Nevertheless, to enhance test discriminatory power it would be possible to increase trial numbers with the inclusion of additional actors displaying varied movements in Phase 1.

The Phase 1 videos displayed actors from different viewpoints providing varying levels of full body, gait and facial movement information. The LTFMT line-ups displayed head-and-shoulders information only. It is possible that some participants concentrated on non-face features in the Phase 1 videos which would have reduced performances. A practice line-up trial with cartoon characters was included prior to all Phase 1 videos. However, including a human hybrid-video style line-up in future may more effectively guide participants to concentrate on relevant cues only and this may improve the results of some. On the other hand, body information may enhance identication over faces displayed alone (e.g. Burton, Wilson, Cowan, & Bruce, 1999; Noyes, Hill, & O'Toole, 2018; Robbins & Coltheart, 2015), although body and face matching performances may not correlate (Noyes et al., 2018). For policing and security it might be advantageous to deploy participants with exceptional face *and* body/gait recognition skills, particularly for roles reviewing CCTV footage, or surveillance and viewing suspects in real life. As such, future research could investigate this with whole body hybrid-video line-ups.

Finally, from a theoretical perspective, only long-term memory of the human face here was tested. Previous research (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.

Conclusions

The results from the three experiments reported here were the first to demonstrate that a substantial minority of participants with outstanding short-term face memory ability, cannot sustain these skills over longer-term retention intervals. This has important implications for policing, as some forces have created specialist super-recogniser units (e.g. Robertson et al., 2016), and if recruitment was based only on the tests commonly used in previous research (e.g., Bate et al., 2018; Russell et al., 2009), the impact on crime detection might be lower than if such a unit contained individuals with the full range of superior skills. A substantial minority of SFRs however did sustain exceptional scores on all tests; and suggestions were made as to how to select staff using various statistical criteria for this type of role. The results also have important theoretical implications as there was no evidence that the shape of the forgetting curve for human faces differed by ability. As such, SFRs' superiority may be primarily based on enhanced early familiarisation to faces, with face memory decay matching the pattern of most of the population.

Context

This research builds on the first author's previous eyewitness identification (e.g. Davis, Valentine, Memon, & Roberts, 2015), and simultaneous face matching research (e.g. Davis & Valentine, 2009), and the first and third authors' super-recognition research (e.g. Belanova et al., 2018). Embedded within policing and law, empirical outcomes have directly linked theory to applied practice. In London, police super-recognisers identified by the research programme have made thousands of suspect identifications (e.g., Davis et al., 2016; 2018), encouraging other international police forces to employ the same recruitment test paradigm (e.g. Munich: Crossland, 2018). The current research demonstrates that for rigorous super-recogniser selection including a long-term face memory test in batteries is essential. This work will likely have long-term impact. Police worldwide are also increasingly deploying face recognition software. Identification accuracy from images is optimised by the fusion of super-recogniser and algorithm decisions (Philips et al., 2018). For legal purposes, final identification decisions are unlikely to ever be made by machines. The research described in the current paper clearly demonstrates that the extremely rare super-recognisers are the most likely to correctly identify a suspect, and to correctly reject innocent suspects regardless of whether decisions require no memory or are delayed for many months.

References

Bate, S., Haslam, C., Jansari, A., & Hodgson, T. L. (2009). Covert face recognition relies on affective valence in congenital prosopagnosia. *Cognitive Neuropsychology, 26*, 391-411. doi: 10.1080/02643290903175004

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. *Quarterly Journal of Experimental Psychology, 70*(2), 193-200. DOI: 10.1080/17470218.2016.1195414

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications,* 3**,** 22, https://doi.org/10.1186/s41235-018-0116-5

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. https://doi.org/10.1016/j.cortex.2018.07.008

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. http://dx.doi.org/10.1016/j.jarmac.2012.02.001

Blunt, M. R., & McAllister, H. A. (2009). Mug shot exposure effects: Does size matter? *Law and Human Behavior, 33*(2), 175-182. DOI: 10.1007/s10979-008-9126-z

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. (2016b). Solving the border control problem: evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS one*, *11*(2), e0148148. https://doi.org/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*. https://doi.org/10.3389/fpsyg.2016.01378

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and “super” face recognition. *The Quarterly Journal of Experimental Psychology*, *70*(2), 201-217. https://doi.org/10.1080/17470218.2016.1161059

Bornstein, B. H., Deffenbacher, K. A., Penrod, S. D., & McGorty, E. K. (2012). Effects of exposure time and cognitive operations on facial identification accuracy: a meta-analysis of two variables associated with initial memory strength. *Psychology, Crime & Law*, *18*(5), 473-490. https://doi.org/10.1080/1068316X.2010.508458

Bretfelean, L-D., & Davis, J.P. (2017). Super-recognition: Long term reliable memory of unfamiliar faces following a ‘fleeting glimpse’. *Undergraduate Prize Winner, British Psychological Society: Cognitive Section Annual Conference,* University of Liverpool, 30 August 2017.

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. https://doi.org/10.1111/1467-9280.00144

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior*, 29(5), 575-604. DOI: 10.1007/s10979-005-5690-7

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

Crawford, J. R., & Garthwaite, P. H. (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex, 48*, 1009-1016. DOI: 10.1016/j.cortex.2011.06.021

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

Crossland, D. (2018). 'Super recognisers' to spot guilty faces at Oktoberfest. *The Times,* 22 September 2018, https://www.thetimes.co.uk/article/super-recognisers-to-spot-guilty-faces-atoktoberfest-06tkfhbjp

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. https://doi.org/10.1002/acp.3260

Davis, J. P., Lander, K., & Jansari, A. (2013). I never forget a face. *The Psychologist, 26,* 726-729.

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. https://doi.org/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*

Davis, J. P., Valentine, T., Memon, A., & Roberts, A. J. (2015). Identification on the street: a field comparison of police street identifications and video line-ups in England. *Psychology, Crime and Law, 1,* 9-27. http://dx.doi.org/10.1080/1068316X.2014.915322

De Haan, E. H., Young, A. W., & Newcombe, F. (1987). Face recognition without awareness. *Cognitive Neuropsychology, 4*, 385– 415. https://doi.org/10.1080/02643298708252045

De Haan, E. H. F., Young, A. W., & Newcombe, F. (1991). Covert and overt recognition in prosopagnosia. *Brain, 114*, 2575– 2591. DOI: 10.1093/brain/114.6.2575

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.

Earles, J. L., Kesten, A. W., Curtayne, E. S., & Perle, J. G. (2008). That's the man who did it, or was it a woman? Actor similarity and binding errors in event memory. *Psychonomic Bulletin & Review, 15*, 1185–89.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (No. 3). University Microfilms.

Edmond, G., & Wortley, N. (2016). Interpreting image evidence: Facial mapping, police familiars and super-recognisers in England and Australia. *Journal of International and Comparative Law, 3*(2), 473-522.

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

Fitzgerald, R., Price, H. L., & Valentine, T. (2018). Eyewitness identification: live, photo, and video lineups. Psychology, Public Policy, and Law. http://dx.doi.org/10.1037/law0000164

Goodsell, C. A., Neuschatz J. S., & Gronlund, S. D. (2009). Effects of mugshot commitment on lineup performance in young and older adults. *Applied Cognitive Psychology, 23*, 788–803.

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

Havard, C., Memon, A., Clifford, B., & Gabbert, F. (2010). A Comparison of Video and Static Photo Lineups with Child and Adolescent Witnesses. *Applied Cognitive Psychology*, *24*(9), 1209-1221. DOI: 10.1002/acp.1645

Hinz, T., & Pezdek, K. (2001). The effect of exposure to multiple lineups on face identification accuracy. *Law and Human Behavior*, *25*, 185–198. http://psycnet.apa.org/doi/10.1023/A:1005697431830.

Jenkins R., Dowsett A.J., & Burton A.M. (2018). How many faces do people know? *Proceedings of the Royal Society (B), 285,* 20181319. http://dx.doi.org/10.1098/rspb.2018.1319

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3-28. http://dx.doi.org/10.1037/0033-2909.114.1.3

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429-442. https://doi.org/10.1080/13506280444000382

Lander, K., & Davies, R. (2007). Exploring the role of characteristic motion when learning new faces. *Quarterly Journal of Experimental Psychology*, *60*(4), 519-526. http://psycnet.apa.org/doi/10.1080/17470210601117559

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

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

McDonald, RP (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum

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

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

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology, 19,* 545-567. http://dx.doi.org/10.1002/acp.1097

Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. *British Journal of Psychology, 94*(3), 339-354. DOI: 10.1348/000712603767876262

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*. DOI: 10.1037/a0030609Vol 18(4), Dec 2012, 361-376

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: using individual data in the interpretation of group results. *Cognitive research: principles and implications*. https://doi.org/10.1186/s41235-018-0117-4

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.

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., … McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology, 70,* 218–233. https://doi.org/10.1080/17470218.2016.1161058

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K. et al. (2018). Face recognition accuracy of forensic examiners, super-recognizers, and face recognition algorithms. *Proceedings of the National Academy of Science (PNAS).* http://www.pnas.org/content/early/2018/05/22/1721355115

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, 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. http://dx.doi.org/10.1371/journal.pone.0150036

Rossion, B., Caldara, R., Seghier, M., Schuller, A. M., Lazeyras, F., & Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus in necessary for normal face processing. *Brain, 126,* 2381-2395. https://doi.org/10.1093/brain/awg241

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

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. http://dx.doi.org/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. http://dx.doi.org/10.1073/pnas.1421881112

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*(2), 139-156. http://dx.doi.org/10.1037/0033-2909.100.2.139

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

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

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*, 99-139. http://dx.doi.org/10.1037/a0021650

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

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352-373. http://dx.doi.org/10.1037/h0020071

Valentine, T., Davis, J. P., Memon, A., & Roberts, A. (2012). Live showups and their influence on a subsequent video lineup. *Applied Cognitive Psychology, 26(1),* 1-23. DOI: 10.1002/acp.1796

Wells, G. L., Smalarz, L, & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition, 4,* 313-317. http://dx.doi.org/10.1016/j.jarmac.2015.08.008

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. https://doi.org/10.1348/000712600161709

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

Wilmer, J. B., Germine, L., Loken, E., Guo, X. M., Chatterjee, G., Nakayama, K., Duchaine, B. (2010). Response to Thomas: is human face recognition ability entirely genetic? *Proceedings of the National Academy of Sciences of the USA*, *107*, E101. DOI: 10.1073/pnas.1004299107

Wixted, J. & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121(2), 262-76*. DOI: 10.1037/a0035940

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.

Supplementary data (for online presentation)

Table S1: Experiment 1 LTFMT8 outcomes by condition for SFRs (SFRs) (n =112) and controls (n = 222)

		SFRs								Controls							
		Hybrid-video				Photo				Hybrid-video				Photo			
		Warning		No warning		Warning		No warning		Warning		No warning		Warning		No warning	
Phase 1		30s	60s	30s	60s	30s	60s	30s	60s	30s	60s	30s	60s	30s	60s	30s	60s
n		35		31		24		22		60		70		55		37	
Target-present																	
Hits	*M*	0.47	0.63	0.49	0.55	0.39	0.29	0.56	0.57	0.40	0.40	0.42	0.44	0.31	0.37	0.50	0.31
	SD	0.27	0.37	0.27	0.33	0.25	0.33	0.37	0.39	0.28	0.37	0.27	0.34	0.19	0.35	0.29	0.38
Foil	*M*	0.29	0.24	0.35	0.32	0.33	0.38	0.32	0.30	0.33	0.33	0.41	0.37	0.36	0.38	0.39	0.49
IDs	*SD*	0.30	0.35	0.37	0.33	0.32	0.34	0.39	0.33	0.40	0.36	0.37	0.37	0.34	0.37	0.31	0.40
Misses	*M*	0.30	0.13	0.19	0.13	0.35	0.33	0.18	0.14	0.35	0.27	0.24	0.19	0.43	0.25	0.16	0.20
	SD	0.30	0.26	0.33	0.22	0.35	0.28	0.25	0.23	0.36	0.33	0.34	0.27	0.33	0.30	0.26	0.30
Target-absent																	
CRs	*M*	0.53	0.56	0.37	0.39	0.46	0.54	0.34	0.34	0.37	0.38	0.32	0.23	0.37	0.35	0.27	0.27
	SD	0.34	0.42	0.41	0.42	0.39	0.33	0.36	0.42	0.34	0.42	0.35	0.35	0.35	0.36	0.33	0.37
Signal detection theory																	
$d^{/}$	*M*	0.12	0.99	-0.45	-0.09	-0.44	-0.13	-0.25	-0.12	-0.76	-0.47	-0.90	-1.06	-1.08	-0.69	-0.86	-1.15
	SD	1.97	2.51	2.15	2.28	2.01	1.84	1.91	2.44	1.86	2.14	1.82	1.85	1.74	1.87	1.54	1.82
C	*M*	0.06	-0.25	-0.32	-0.43	0.05	0.24	-0.55	-0.61	-0.18	-0.26	-0.31	-0.62	<0.01	-0.27	-0.54	-0.40
	SD	0.75	0.88	0.91	0.95	0.85	0.71	1.09	0.95	0.85	1.01	0.88	0.87	0.72	0.89	0.97	1.00
Confidence																	
C	*M*	60.7	64.9	58.9	64.2	58.3	59.3	66.7	67.3	57.0	57.7	49.8	52.0	56.4	58.4	45.8	44.5
	SD	19.2	21.0	22.4	20.9	23.3	28.8	16.3	18.6	21.8	23.5	18.4	17.9	20.5	19.5	23.0	23.2

Table S2: Results of 2 (group: SFRs (SFRs), controls) x 2 (Phase 1 display time (DT): 30s, 60s) x 2 (Phase 2 line-up media (LM): hybrid-video, photo) x 2 (Warning (Warn): warning, no warning) ANOVAs conducted on LTFMT8 outcomes in Experiment 1. Values of F and η^2 are reported. Post hoc test or simple effects analyses employed t-tests (and Cohen's d) or F (and η^2)

df	Hits		Foil IDs		Misses		CRs		Sensitivity (d$^{/}$)		Criterion (C)		Confidence	
(1, 326)	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
Main effects														
Group: GP	12.64 *	.037	3.95 *	.012	2.23	.007	13.04 *	.038	22.87 *	.066	1.35	.004	17.70 *	.051
Display time: DT	<1	<.001	<1	<.001	9.68 *	.029	<1	<.001	2.34	.007	2.28	.007	5.50 *	.017
Lineup media LM	5.10 *	.015	1.17	.004	1.28	.004	<1	.002	2.31	.007	<1	<.001	<1	<.001
Warning: W	6.83 *	.021	1.20	.004	19.81 *	.057	15.15 *	.044	3.05	.009	23.51 *	.067	1.56	.005
Two-way interactions														
GP x DT	1.32	.004	<1	<.001	<1	<.001	<1	.003	1.39	.004	<1	<.001	1.48	.005
GP x LM	<1	.001	<1	<.001	1.44	.004	<1	.001	<1	.001	<1	<.001	<1	<.001
GP x W	<1	.003	<1	.002	<1	<.001	<1	.003	<1	<.001	1.69	.005	7.66 *	.023
SFR													*F* <1	*d* .003
Controls													*F* 11.77	*d* .035
DT x LM	5.19 *	.019	1.41	.004	<1	.002	<1	.001	<1	<.001	1.49	.007	2.58	.008
30s	*t* <1	*d* .007												
60s	*t* 2.65 *	*d* .029												
DT x W	1.37	.004	<1	<.001	3.11	.009	<1	.002	2.14	.007	<1	<.001	<1	<.001
LM x W	6.28 *	.019	<1	.002	3.08	.009	<1	<.001	2.83	.009	2.65	.008	<1	<.001
Photo	*F* 12.10 *	η^2 .036												
Video	*F* <1	η^2 .002												
Three-way interaction (only this 3-way was significant)														
GP x LM x W	3.88 *^{A}	.012	<1	<.001	<1	<.001	<1	<.001	<1	.001	1.13	.003	2.64	.008
Four-way interaction														
4-way	6.03 *^{A}	.018	<1	<.001	3.05	.009	<1	.002	<1	.001	4.37 *^{A}	.013	1.02	.003

* $p < .05$

^{A} Interaction effects reported in text

Table S3: Results of 2 (group: SFR, controls) x 2 (Decision Commitment: identical, different) x 2 (Experiment 2 Decision: correct target, incorrect foil) ANOVAs conducted on LTFMT8 outcomes in Experiment 2 if a selection was made from a line-up in Experiment 1 and 2.

df		Hits		
(1, 155)		*F*		η^2
Main effects				
Group		2.43		.015
Decision Commitment		6.23 *		.039
Experiment 2 Accuracy		31.83 *		.170
Two-way interactions				
Group x Commitment		7.85 *		.048
Identical	*t*	1.51	*d*	.25
Different	*t*	3.07 *	*d*	.52
Group x Accuracy		30.80 *		.166
Correct targets	*t*	4.81 *	*d*	.77
Incorrect Foils	*t*	-4.59 *	*d*	.68
Commitment x Accuracy		14.81 *		.087
Identical	*t*	2.84 *	*d*	.35
Different	*t*	7.91 *	*d*	.91
Three-way-interaction				
Group x Commitment x Accuracy		1.36		.009

* $p < .05$

Table S4: Mean Experiment 2 hits, foil IDs, and misses to target-present line-ups that had been correctly rejected when target-absent in Experiment 1

	Superior-Face Recognisers		Control		Total	
	n = 49		*n* = 73		*n* = 122	
	M	*SD*	*M*	*SD*	*M*	*SD*
Hits	0.45	0.71	0.22	0.48	0.31	0.59
Foil IDs	1.24	0.88	1.16	1.15	1.20	1.05
Misses	0.65	0.78	0.55	0.69	0.59	0.72
Total	2.35	0.95	1.93	0.93		

Table S5: Demographic data and group inclusion criteria, and results for t-tests comparing self-belief in ability (1-5), CFMT+, STFMT and LTFMT10 outcomes in Experiment 3 (regardless of delay)

	All		SFRs		Controls					
n	597		84		103		*df*	*t*	*d*	*p*
Age	16-74 yrs.		19-64 yrs.		16-74 yrs.					
	M	*SD*	*M*	*SD*	*M*	*SD*				
	35.3	11.6	33.5	9.7	35.7	13.8	181.5	-1.29	0.18	.200
Gender										
Male	241 (40.4%)		33 (39.3%)		55 (53.4%)			[A]		
Ethnicity										
White	505 (84.3%)		69 (83.1%)		86 (83.5%)			[B]		
			SFR and control group inclusion criteria							
CFMT+			95-102		58-83					
STFMT d$^{/}$			> 1.8147		0.8932-1.8147					
	M	*SD*	*M*	*SD*	*M*	*SD*				
Self-belief	3.80	0.90	3.99	0.94	3.76	0.82	184	1.73	0.26	.085
CFMT+	86.6	9.52	97.1	1.90	75.3	6.27	124.4	33.33	4.71	<.001
STFMT										
Hits	0.77	0.13	0.87	0.08	0.72	0.11	183.9	10.32	1.56	<.001
CRs	0.82	0.11	0.90	0.07	0.77	0.10	183.5	10.76	1.51	<.001
d$^{/}$	1.84	0.66	2.60	0.53	1.38	0.25	112.3	19.30	2.94	<.001
C	0.10	0.34	0.07	0.31	0.09	0.34	185	-0.31	0.06	>.2
GFMT										
Hits	0.93	0.08	0.97	0.05	0.89	0.09	167.5	6.82	1.10	<.001
CRs	0.93	0.08	0.96	0.06	0.88	0.11	158.4	6.51	0.90	<.001
d$^{/}$	3.12	0.66	3.48	0.49	2.64	0.68	182.7	9.91	1.42	<.001
C	<0.01	0.27	-0.03	0.21	-0.03	0.34	174.7	0.06	<0.01	>.2
LTFMT10										
Delay	24.1	23.9	27.4	28.9	24.6	23.7	185	0.72	0.11	>.2
Hits	0.41	0.23	0.52	0.22	0.33	0.19	185	6.25	0.92	<.001
Foil IDs	0.44	0.23	0.32	0.22	0.51	0.21	185	-6.02	0.88	<.001
Misses	0.16	0.21	0.16	0.21	0.16	0.20	185	0.08	<0.01	>.2
Conf	57.4	21.5	61.5	21.9	57.0	22.0	185	1.38	0.21	>.2

[A] Gender (Male = 1 *vs.* female = 0) x group, χ^2 (1, 187) = 3.70, *p* = .055, Cramer's V = .141.

[B] Ethnicity (White = 1 *vs.* other ethnicity = 0) x group, χ^2 (1, 179) < 1.

[C] Conf = Confidence

Table S6: Pearson's correlation coefficients between outcome measures in Experiments 1 to 3 (n varies between and within experiments due to missing data caused by participant dropout or lack of responses).

	GFMT	CFMT+	STFMT	Experiment 1 LTFMT8			Experiment 2 LTFMT8			Experiment 3 LTFMT10			*M*
	/40	/102	d$^{/}$	Delay	d$^{/}$	Confidence	Delay	Hits	Confidence	Delay	Hits	Confidence	(*SD*)
n	4437	4444	4444	1091	1091	1091	539	539	539	595	594	595	3.72
Self-Belief	0.19 *	0.25 *	0.17 *	0.01	0.07	0.20 *	-0.06	0.17 *	0.19 *	0.07	0.03	0.13 *	(0.88)
n		4451	4451	1085	1085	1085	535	535	535	597	596	597	36.80
GFMT		0.47 *	0.32 *	-0.07	0.17 *	0.11 *	0.03	0.19 *	0.05	0.03	0.21 *	0.05	(2.68)
n			4458	1091	1091	1091	539	539	539	597	596	597	83.66
CFMT			0.41 *	-0.02	0.18 *	0.16 *	0.01	0.23 *	0.10	0.02	0.22 *	0.07	(10.80)
n				1091	1091	1091	539	539	539	597	596	597	1.77
STFMT d$^{/}$				-0.02	0.21 *	0.16 *	-0.07	0.26 *	0.11	-0.03	0.23 *	0.10	(0.67)
Experiment 1 LTFMT8													
n					1091	1091	539	539	539				8.97
Delay (days)					-0.04	-0.05	-0.13 *	-0.03	0.07				(6.03)
n						1091	539	539	539				-0.56
Sensitivity (d$^{/}$)						0.27 *	-0.04	0.26 *	0.13				(1.49)
n							539	539	539				57.4
Confidence							-0.04	0.22 *	0.57 *				(20.1)
Experiment 2 LTFMT8													
n								539	539				170.7
Delay (days)								0.03	-0.13				(40.3)
n									539				0.24
Hit rates									0.15 *				(0.17)
n													43.9
Confidence													(19.1)
Experiment 3 LTFMT10													
n											596	597	24.1
Delay (days)											-0.26 *	-0.07	(29.9)
n												596	0.41
Hit rates												0.31 *	(0.23)
n													57.4
Confidence													(21.5)

* To control for Type-I errors and for trivial findings with large numbers of participants, α = .001.

Table S7. Multiple regression analyses for variables predicting Experiment 1's LTFMT8 sensitivity (d$^{/}$) R^2 = 0.11, F(6, 1078) = 23.05, p < .001

	Unstandardised Coefficients	Standardised Coefficients					
	B	SEM	β	*t*	*p*	95% CI	
(Constant)	-4.74	0.72		-6.58	<.001	-6.15	-3.32
Self-belief	-0.03	0.06	-0.02	-0.53	.602	-0.14	0.08
GFMT	0.05	0.02	0.08	2.60	.009	0.01	0.09
CFMT+	0.01	0.01	0.06	1.87	.062	<0.01	0.02
STFMT d$^{/}$	0.29	0.07	0.13	4.08	<.001	0.15	0.43
Delay	-0.01	0.01	-0.02	-0.74	.459	-0.02	0.01
Confidence	0.02	<0.01	0.23	7.76	<.001	0.01	0.02

Table S8. Multiple regression analyses for variables predicting Experiment 2's repeated LTFMT8 hit rates, $R^2 = 0.14$, $F(9, 525) = 10.33$, $p < .001$

	Unstandardised Coefficients	Standardised Coefficients					
	B	SEM	β	*t*	*p*	95% CI	
(Constant)	-0.27	0.13		-2.14	.033	-0.52	-0.02
Self-belief	0.01	0.01	0.07	1.58	.115	<-0.01	0.03
GFMT	<0.01	<0.01	0.05	1.16	.248	<-0.01	0.01
CFMT+	<0.01	<0.01	0.09	1.77	.077	<0.01	<0.01
STFMT d$^{/}$	0.04	0.01	0.15	3.20	.001	0.02	0.06
Experiment 1							
Delay	<0.01	<0.01	-0.01	-0.12	.908	<-0.01	<0.01
d$^{/}$	0.02	<0.01	0.17	3.84	<.001	0.01	0.03
Confidence	<0.01	<0.01	0.09	1.79	.075	<0.01	<0.01
Experiment 2							
Delay	<0.01	<0.01	0.06	1.42	.155	<0.01	<0.01
Confidence	<0.01	<0.01	0.04	0.80	.426	<-0.01	<0.01

Table S9. Multiple regression analyses for variables predicting Experiment 3's LTFMT10 hit rates, $R^2 = 0.21$, $F(6, 587) = 27.40$, $p < .001$.

	Unstandardised Coefficients	Standardised Coefficients					
	B	SEM	β	*t*	*p*	95% CI	
(Constant)	-4.02	1.34		-3.01	.003	-6.65	-1.40
Self-belief	-0.09	1.00	-0.03	-0.87	.384	-0.28	0.11
GFMT	0.12	0.04	0.12	2.90	.004	0.04	0.19
CFMT+	0.03	0.01	0.11	2.45	.015	0.01	0.05
STFMT d$^{/}$	0.43	0.14	0.12	3.10	.002	0.16	0.71
Delay	-0.02	<0.01	-0.24	-6.48	<.001	-0.03	-0.02
Confidence	0.03	<0.01	0.27	7.16	<.001	0.02	0.04

Table S10: Self-assessments of ability by super-face-recognisers (n = 422, M = 4.1, SD = 0.8)

	Well above Average	Above Average	Average	Below Average	Well below Average
n	137	211	59	1	4