Identification from CCTV: Assessing police super-recogniser ability to spot faces in a crowd and susceptibility to change blindness

Josh P Davis[1], Charlotte Forrest[1], Felicia Treml[1], and Ashok Jansari [2]

[1] Department of Psychology, Social Work and Counselling, University of Greenwich, London, UK

[2] Department of Psychology, Goldsmiths, University of London, London, UK

Acknowledgements: This project was 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 wish to thank Paul Smith and Mick Neville of the Metropolitan Police Service for organising police recruitment, and for gaining permission from landowners – who we also thank - for filming the Spot the Face in a Crowd videos; Diandra Bretfelean, Monika Durova, Hava Sokoli, Donata Andriuskeviciute for assisting with administration; and Andreea Maigut, Oliwia Willetts, Elena Mut, Meike Imberg, Alice Cox, Aurora Trentin, Leandra Sinischali, Robbie Reid, Yana Mihaylova, Maisie Van-Velsen, Tasnim Fakira, Maryam Halaoui, Bankole Osibote, Remi Fletcher for participant testing. We would also like to thank three anonymous reviewers for their extremely helpful comments on a previous version of this manuscript.

Correspondence to:

Dr Josh P Davis
Reader in Applied Psychology
Department of Psychology, Social Work and Counselling
University of Greenwich
Avery Hill
London, SE9 2UG, UK
j.p.davis@greenwich.ac.uk

Word count (excluding titles, abstract, tables, figures, and references): 9,283

This study was approved by the University of Greenwich Research Ethics Committee.

Abstract

Police worldwide regularly review CCTV evidence in investigations. This research found London *police experts* who work in a full-time 'Super-Recogniser Unit' and front line *police identifiers* regularly making suspect identifications from CCTV, possessed superior unfamiliar face recognition ability, and, with higher levels of confidence, outperformed controls at locating actors in a bespoke *Spot the Face in a Crowd Test* (SFCT). Police were also less susceptible to change blindness errors, and possessed higher levels of conscientiousness, and lower levels of neuroticism and openness. Controls who took part in SFCT actor familiarisation training outperformed untrained controls, suggesting this exercise might enhance identification of persons of interest in real investigations. This research supports an accumulating body of evidence demonstrating that international police forces may benefit from deploying officers with superior face recognition ability to roles such as CCTV review, as these officers may be the most likely to identify persons of interest.

Key words: Super-recogniser, face recognition, CCTV, face matching, change blindness,

Introduction

Police worldwide regularly review CCTV evidence in investigations. Officers construct event timelines, search for persons of interest (e.g., missing persons, victims, witnesses), and identify suspects. Reviews can be time consuming, particularly during major operations (Warm, Parasuraman, & Matthews, 2008). Following the 2011 London Riots over 200,000 hours of footage was searched and 4,000 suspects identified (BBC News, 2011). Reviewers must sustain attention and keep in mind memorial images of targets. Research has examined criminal behaviour anticipation, threat detection (e.g., Gelernter, 2013; Troscianko et al., 2004), and task disengagement (Donald & Donald, 2014) during CCTV viewing (for a review see Hillstrom, Hope, & Nee, 2008). Vigilance tasks are associated with high mental workload (e.g., Warm et al., 2008), and are influenced by experience (Biggs, Cain, Clar, Darling, & Mitroff, 2013), fatigue (e.g., Wickens & McCarley, 2008), and stress (e.g. Sawin & Scerbo, 1995). Automated search aids (Warm, Dember, & Hancock, 1996), practice (McCarley, Kramer, Wickens, Vidoni, & Boot, 2004; Uenking, 2000) and caffeine (Temple et al., 2007) enhance vigilance in other domains (e.g., airport baggage screening, airline pilots, and radiology). However, considering homeland security and policing implications, research examining whether CCTV search can be optimised by deploying staff with specific competencies has been limited. The current research therefore aimed to examine whether human (e.g., face recognition ability, personality, perceived workload, susceptibility to change blindness) or operational factors (e.g. operational experience, face familiarisation training, target search numbers) influence ability to locate persons of interest in video.

Individual differences and superior face recognition ability

There are large individual differences in face recognition ability, with developmental prosopagnosics (DPs) and super-recognisers (SRs) at the extremes. Substantial research has investigated DPs' poor abilities, whereas few studies have examined SRs' superiority (e.g., Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Davis, Lander, Evans, & Jansari, 2016; Russell, Duchaine, & Nakayama, 2009, for a review see Noyes, Phillips, & O'Toole, 2017). For SR group inclusion, most researchers use self-reports of exceptional ability, and scores in the top 2% of the population on the 102-trial *Cambridge Face Memory Test: Extended* (CFMT+) (Russell et al., 2009). However, self-reports may be unreliable (Bobak, Pampoulov, & Bate, 2016), and different CFMT+ thresholds have been employed (e.g. 90/102; Bobak, Bennetts et al., 2016; 95/102; Bobak, Pampoulov, & Bate, 2016). Moreover, some high-threshold achieving SRs (CFMT+ $\leq$ 95) perform poorly at simultaneous face matching (e.g., Bobak, Hancock, & Bate, 2016; Davis et al., 2016), suggesting SR may be an umbrella term for heterogeneous constructs. Although the CFMT+ appears to be a good marker of SR ability, Noyes et al. (2017) suggest additional tests are required for reliable SR 'diagnosis'. However, multiple tests can induce fatigue, a problem when assessing exceptional ability, and for the current research, participants completed the CFMT+ only, and data are reported of those achieving the two SR thresholds described above.

Face processing and law enforcement

Exceptional face processing ability has policing implications. London's Metropolitan Police Service (MPS) instituted a 'SR pool' of front-line officers (n ~ 150), after they identified a substantial minority of 2011 London rioters from CCTV (Davis, Lander, & Jansari, 2013; Evison, 2014). Others joined later after making multiple identifications from the MPS' *Caught on Camera* wanted suspect website. As not all have been empirically

tested, here they are described as *police identifiers*. Their successes contrast with most of the remaining 48,000 strong MPS workforce who rarely make identifications, although they may not be familiar with any captured on camera, or don't view images in the first place. Police identifiers are given regular time to view often locally filmed suspect images. Most identify familiar suspects, although familiarity, and numbers viewed and identified substantially vary. Ground truth of guilt cannot always be determined, but between May 2013 and Oct 2015, police identifiers (n = 143) made over 9,500 identifications. Over 50% of suspects were charged to appear in court once additional evidence was secured.

Some police identifiers have been deployed to review CCTV footage to identify persons of interest, and to CCTV control rooms or strategic locations at pop concerts (Davis et al., 2013), or events such as London's Notting Hill Carnival with crowds of more than a million (Venkataramanan, 2015). They attempt to spot unfamiliar suspects whose photographed faces they have committed to memory. A few (n = 36) have been tested on familiar and unfamiliar face recognition and face matching tests (Davis et al., 2016). Many matched the performances of 10 high (≤ 95) CFMT+ threshold SRs. Virtually all the remainder outperformed the mean scores of demographically matched controls (n = 143).

In 2015, the MPS established a full–time police 'Super-Recogniser Unit' (n = 7) (hereafter *police experts*), dedicated to CCTV footage review. No information as to exactly how police experts are selected or deployed is available, except all were police identifiers. Media interviews suggest that experts regularly recognise and match unfamiliar suspects across images from different crime scenes (e.g. Manzoor, 2016). These reports replicate SR anecdotes, who claim that familiarisation with unfamiliar faces is easy (e.g. Russell et al., 2009). This, if supported, is impressive as face recognition and matching (Burton, Wilson, Cowan, & Bruce, 1999; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016), and searching for faces in crowds (Ito & Sakurai, 2014) is more accurate with familiar than unfamiliar

faces. Unsurprisingly, some police experts (n = 4) performed 'well above normal' on empirical tests of unfamiliar and familiar face matching (Robertson et al., 2016).

CCTV review, attention, vigilance and personality

To address factors that may contribute to CCTV review, and searches for persons of interest, in the research reported here, police experts, police identifiers and non-police controls completed the CFMT+, before searching for unfamiliar target-actors in a novel video review *Spot the Face in a Crowd Test* (SFCT). It was designed to examine whether such tests might assist deployment decisions to review roles, and whether SFCT performance was mediated by experience, and/or face recognition ability. The SFCT draws on memory, as well as perception, as participants compare actor photos with the crowds on video, either when moving or by pausing to view individual frames. Nevertheless, unfamiliar face matching can be unreliable (e.g., Bruce et al., 1999; for a review see Davis & Valentine, 2015), particularly if, as with the SFCT, targets have changed appearance – as often encountered by police. Distractor frequency (Singh, Tiwari, & Singh, 2007; Wickens, Gempler, & Morphew, 2000); and target regularity (Wolfe, Horiwitz, & Kenner, 2005), and numbers (Tickner & Poulton, 1975) can influence vigilance tasks. In the SFCT, participants were assigned two, four, or eight target-actors to search for allowing an examination of this factor on performance.

Bruce, Henderson, Newman, and Burton (2001) found that simultaneous face matching was enhanced if participants discussed in pairs the faces' perceived personalities in advance. This procedure, possibly promoting deeper social processing, and drawing attention to effective identification cues (see also McGugin, Tanaka, Lebrecht, Tarr, & Gauthier, 2011) was adopted here. In pairs, some controls engaged in pre-SFCT actor *familiarisation* training, and were expected to outperform *untrained* controls.

Personality may also impact CCTV review. Extraversion may *positively* correlate with face memory (Lander & Poyarekar, 2015; Li et al., 2010; Megreya & Bindemann, 2013); whereas, extraversion (Davies & Parasuraman, 1982), and neuroticism (Derakshan & Eysenck, 2009; Sadeh & Bredemeier, 2011) *negatively* correlate with visual search and vigilance performance. Furthermore, the emotional *Face in a Crowd* paradigm suggests that when displayed in arrays, threatening or angry faces grab attention more than neutral faces. Indeed, Damjanovic, Pinkham, Clarke, and Phillips (2014) found that in comparison to less experienced police, experienced riot police detect threating images more effectively. This highlights experiential factor influence, although SFCT actors were asked to act 'naturally'. To address this contradictory research in the context of CCTV review, participants here completed measures of personality (*International Personality Item Pool Representation of the NEO PI-R: IPIP-NEO*: Goldberg, 1999), perceived SFCT workload (*National Aeronautics and Space Administration Task Load Index: NASA-TLX*: Hart & Staveland, 1988), and a change blindness susceptibility test (Smart, Berry, & Rodriguez, 2014).

Change blindness and video footage review

CCTV review requires sustained concentration. However, *inattentional blindness* paradigms demonstrate that focussed attention can blind viewers to anomalies. For instance, when counting passes made by a basketball team, many participants fail to notice a man in a gorilla costume (Simons & Chabris, 1999). Moreover, *change blindness* paradigms demonstrate that environmental changes, including people, are often missed (e.g., Simons & Levin, 1998; Smart et al., 2014, for a review see Gibbs, Davies, & Chou, 2016). For instance, Smart et al. (2014) found 44.6% of participants failed to notice that a police-stopped driver was replaced by a second driver after briefly walking off camera. Those missing the switch

made more line-up identification errors later. Change blindness might influence CCTV review if persons of interest are tracked through footage from different cameras, with gaps in the field of view. However, Beck, Martin, Smitherman, and Gashen (2013) found that laypersons make more change blindness errors than experts (in this case veterinarian radiologists). More effective encoding of task-relevant information reduces susceptibility. This might imply police would be immune during police-relevant scenarios. However, Smart et al. (2014) found police and students equally missed the driver switch, although they did not test face recognition ability. Excellent skills might provide immunity from person change blindness. To test this, the same change blindness test was employed here, with dependent variables being driver change detection or not, and subsequent line-up accuracy. It was completed shortly after the SFCT when fatigue and change blindness susceptibility might be expected to be highest. An alternative prediction was that having become acclimatised to video analysis, participants, particularly those with superior face recognition ability, and/or CCTV review experience, might be less susceptible to change blindness, and be more likely to identify the drivers from the line-ups.

The current research

.

In summary, the current research employed the CFMT+ (Russell et al., 2009), the SFCT, the Change Blindness Test (Smart et al., 2014) and follow-up target-absent (n = 2) and target-present line-ups (n = 2), the latter containing the two switched drivers; as well as the *NASA-TLX* (Hart & Staveland, 1988) and the *IPIP-NEO* (Goldberg, 1999). As is common in SR research (e.g. Bobak, Bennetts et al., 2016), to complement between-group analyses, neuropsychological-style individual analyses were conducted (e.g. Crawford, Garthwaite, & Porter, 2010). The performances of all participants meeting high SR CFMT+ threshold (≤

95), were compared with the mean scores of police identifiers not meeting SR criteria (i.e. CFMT+ < 90), as this group were most likely to provide a similarly motivated control group (for a discussion of this issue see Noyes et al., 2017).

A number of hypotheses were derived. As most police experts possess empirically tested superior face processing ability (Robertson et al., 2016), work full time at CCTV review and have made multiple unfamiliar suspect identifications, they were expected to outperform the police identifiers and controls at the CFMT+, the SFCT, and the Change Blindness Test. As many police identifiers also possess superior face recognition ability (Davis et al., 2016), they were predicted to outperform controls. Consistent with previous research (Tickner & Poulton, 1975), there was additionally expected to be a negative relationship between target-actor search quantity and SFCT accuracy, with familiarised controls predicted to outperform untrained controls (see Bruce et al., 2001). As this was exploratory research, no predictions were made in relation to perceived workload, or personality, and SFCT performance.

Method

Design

A correlational component examined the relationships between performances on three cognitive tests: - *CFMT+* (Russell et al., 2009), *SFCT*, and *Change Blindness* (Smart et al., 2014), the latter operationalised by whether participants noted the driver change and identified them later from line-ups; as well as the 'big five' personality factors (*IPIP-NEO*, Goldberg, 1999), and for the SFCT, a perceived workload measure (*NASA-TLX*, Hart & Staveland, 1988). An independent-measures component compared performances of MPS SR

Unit members (police experts); MPS front line SRs (police identifiers); and non-police controls. Individual analyses also compared performances of all participants meeting high SR threshold (CFMT+ ≤ 95) against the mean of non-SR police identifiers.

Additional factors for controls were *training*, or assignment to the pre-SFCT actor familiarisation exercise or not (*familiarised* vs. *untrained*), and for controls and police identifiers, *actor-number* (participants searched for *two*, *four* or *eight* actors). Due to low numbers, police experts always searched for eight actors. The primary SFCT dependent variables were correct actor identifications (hits), and confidence; false positives of bystanders (FPs), and confidence; and correct rejections of clips 'empty' of actors.

Participants

Police experts were the entire membership (at the time) of the full-time MPS SR unit (*n* = 7, males = 5; white-Caucasian = 100%, age = 28-48 years; *M* = 37.6, *SD* = 7.1). *Police identifiers* were front line members of the cross-London SR pool (*n* = 92; males = 71; White-Caucasian = 87.9%, age = 20-52 years; *M* = 34.3, *SD* = 6.5). The authors are not party to private MPS decisions as to how group members were appointed, and this cannot be reported. Data of all suspect identifications from May 2013-Oct 2015 were supplied by the MPS for the police experts (*n* = 7, Max = 296, *M* = 153, *SD* = 98), and for police identifiers who were pool members in May 2013 (*n* = 46; Max = 481, *M* = 58, *SD* = 86). These suspect identification rate data were incomplete for the remaining police identifiers who joined later (*n* = 46).[1] Additional police identifiers did not participate (*n* ~ 70).

[1] There were no significant correlations between suspect identification rate data of police experts and police identifiers for whom full data were available (*n* = 53), and any CFMT+, SFCT, or Change Blindness Test measures (*p* > .2). There were also no differences between police identifiers with or without suspect identification data on any of these tests (*p* > .1).

Controls were University of Greenwich staff and students (n = 152; males = 37; White-Caucasian = 69.1%; age = 18-66 years; M = 26.7, SD = 11.3). Most received small financial compensation, although some students could claim participation points as part of research methods training. No record was kept of who claimed points or compensation. However, allocation to conditions was randomised.

Chi-squared tests comparing police experts, police identifiers, and familiarised and untrained controls on gender, $\chi^2(3, 251) = 67.06$, $p < .001$, $\Phi = .517$, and ethnicity proportions (white vs. other), $\chi^2(3, 250) = 12.55$, $p = .006$, $\Phi = .224$, and an ANOVA on mean age, $F(3, 244) = 16.67$, $p < .001$, $\eta^2 = .170$, revealed significant differences. Police were mainly male, white, and over 30-years. Controls were mainly female, under 25-years, with a higher proportion of non-whites.

Measures

Cambridge Face Memory Test: Extended (CFMT+; Russell et al., 2009). In the original 72-trial standardised CFMT (Duchaine & Nakayama, 2006), participants are familiarised with, and recognise six unfamiliar hairstyle-obscured faces from increasingly difficult arrays. Targets are displayed from different viewpoints, while expressions change. In the extended 102-trial version (CFMT+), the additional 30 trials are degraded with visual noise, neckline and hairstyles are sometimes depicted, and distractors recur more regularly.

Spotting a Face in a Crowd Test (SFCT): This bespoke test was designed with MPS assistance in selecting photos, videos, and in facilitating London tourist location permissions. The final test consists of an 18 minute 21 second video split into 11 clips (labelled A to K; see Table 3). Clips were recorded using a fixed Canon Fine Pix 5200HD camera on a tripod

and edited using Windows Live Movie Maker. They were taken from a gantry above ground level, a tripod from head height, or on raised ground. Eight white-Caucasian target-actors (*a-h*; aged 19-54, female = 7) were filmed briefly walking through the field of view. The clips contained two (*n* = 2 clips), one (*n* = 7), or zero actors (*n* = 2). Actors appeared in one (*n* = 5 actors) or two (*n* = 3) clips. Actor and bystander distance to camera, and bystander numbers varied substantially. [2]

Four photos of each actor were displayed in colour on 210 × 297 mm paper (see Figure 1). These were self-selected by the actors, instructed to provide those that their families might give to the police if reported missing (i.e. from social media – taken within one-year), and to ensure appearance varied (e.g., clothing, hairstyle). Some actors were inadvertently depicted in one photo and the video wearing the same/similar item of clothing.

Participants booked lab times. Paired slots were available. They were instructed that they would review video footage as though searching for 'missing persons'. Apart from police experts, - always allocated eight actors, participants were provided with the four photographs of two, four or eight actors. Actor assignment for the two- and four-actor conditions was randomised, although within each experimental condition, each actor was searched for an equal number of times to ensure no bias from some being harder to spot.

Approximately half the controls who signed up for a paired slot were randomly allocated to pre-test actor *familiarisation* training.[3] They were advised that face learning is enhanced by deep social processing, and to briefly discuss each actor's perceived personality from their photos by verbally agreeing a rating from 1-10 to a series of questions asking, how *adventurous, good-natured; intelligent; cheerful; creative; stylish; responsible; friendly;*

[2] Videos were taken for the LASIE project (www.lasie-project.eu). These tourist sites regularly charge film crews to take footage depicting background members of the public, and similar warning signs about filming were in place. Publication of location images was not permitted.

[3] Solo controls were assigned to the untrained control group. Numbers in this group assigned to the eight-actor condition were higher than the two- and four-actor groups to ensure that target-actor images were equally distributed.

perfectionist; honest; insightful; competitive; confident; ambitious; calm; sociable; tidy; independent; determined; conservative; caring; outspoken does the actor appear in these photos?" Participants were encouraged to resolve disagreements, and assess, a) How old is the actor? b) Is their face wide or narrow? c) What are their most distinctive facial features? Describe the d) eyes e) mouth f) face shape with your partner. This took up to 20 min. No responses were recorded. *Untrained* participants (controls and all police) started immediately.

All participants completed the SFCT individually in separate booths. They were instructed to use the video player controls (e.g., rewind, pause), and warned that 'some clips depict no actors; whereas others depict one, or more than one'. They could simultaneously compare photos and videos. If they believed they located an actor they paused the video and wrote details on a paper form, which provided space for up to four identification decisions for each clip. For each identification, participants reported the video time from the display, a brief actor description, and decision confidence (3: 'not at all confident', 2: 'possible identification', 1: 'highly confident').[4] There were no time limits (Range = 19 – 150 min).

Responses were analysed by clip (see Table 3). Two researchers independently coded each response by checking if video display time, actor letter, or actor description (clothing descriptions were most informative) matched that of actors or bystanders appearing at that point. They entered responses on separate databases, and once merged, inconsistencies were checked by a third. On the rare occasions the third researcher could not decide ($n < 10$), the first author adjudicated in favour of a correct identification. All adjudicated identifications were made by controls, as police tended to write clearer descriptions.

Responses were coded as hits (correct actor identification), false positives (FP) (incorrect bystander identification), and correct rejections (CR) or correctly making no

[4] Note: Confidence scores were reversed for analyses, so that a high rating (3) = high confidence.

identification to a clip empty of actors.[5] Equal participant numbers within each group (police expert, police identifier, controls) (n = 151) searched for each actor. For those assigned to search for eight actors, the maximum (max) possible hit rate = 11; CR rate = 2. However, actor appearances across clips varied, meaning that for two-actor groups, max hits = 2-4 (M = 2.75, SD = 0.61); CR = 7-9 (M = 8.33, SD = 0.60). For four-actor groups, max hits = 4-7 (M = 5.50; SD = 0.90); CR = 5-7 (M = 5.98; SD = 0.67).

NASA-TLX Load Index (Hart & Staveland, 1988): This five-factor measure evaluated SFCT subjective workload.[6] *Mental Demand* referred to the perceived mental and perceptual activity; *Temporal Demand* to the speed of task requirements; *Effort* to how hard participants perceived the task; *Frustration Level* to stress; and *Overall Performance* to final sense of achievement. Participants read factor descriptions and made 10 pair-wise comparisons as to which of the pair they believed contributed most to the workload. This took less than 5 min.

IPIP-NEO PI-R^{TM}; Goldberg, 1999): The compressed 30-item version of the original 300-item IPIP-NEO inventory measures five domains: neuroticism, extraversion, conscientiousness, openness and agreeableness. Participants self-rated each item on a five-point scale ranging from very accurate to very inaccurate. It took approximately 5 min. The test norms are reported in Table 6.

Change Blindness Test (Smart et al., 2014): Procedures replicated the original research. The 2 min 44 sec video shows a traffic stop from a US police vehicle dashboard. At

[5] The SFCT performance terminology to similar to signal detection theory (SDT) often used in face recognition research. However, calculation of sensitivity or response bias was not possible, as although SFCT hit rates are SDT comparable, FPs of the varying bystander numbers, or CRs of empty clips, are not SDT analogous.

[6] One irrelevant *physical demand NASA-TLX Load Index* scale was excluded. Participants usually provide a rating to each scale. However, due to tine constraints, only the pairwise comparison element was retained.

the start, a vehicle is parked in front of the camera, with an officer standing between the two vehicles. The white-Caucasian male driver (Driver 1) wearing a hat and a *white* t-shirt under a hoodie gets out of his vehicle. Approximately halfway through the video, the driver and police officer step out of view and when they return, Driver 1 (18 years, 188 cm, 56.7 kg) is replaced by Driver 2 (20 years, 175 cm, 61.0 kg) wearing a *green* t-shirt under a similar hoodie. The police officer hands Driver 2 a document, who then returns to 'his' vehicle.

Shortly afterwards, to assess change detection, participants answered two open-ended questions requesting descriptions of the video on Qualtrics online survey software.[7] Ten further four-multiple-choice items measured central and peripheral event memory (e.g., driver trousers colour; number of background vehicles). Participants then viewed a series of four eight-person simultaneous photo line-ups, each arranged in a 2 x 4 array and attempted to identify 'anyone they recognised'. There were no time limits. From 60cm, the visual angle of line-up member presentation was approximately 5.25 by 4.30. The two drivers always appeared in the 2nd and 3rd line-ups in positions 2 and 5 (target-present). Line-ups 1 and 4 were target-absent. Confidence ratings to each decision were collected (1 = not at all confident to 7 = absolutely confident). This test took approximately 10 min.

Two researchers independently rated the open-ended change detection responses on four categories. Was the driver change reported? Was the clothing change reported? Were both changes reported? Were no changes reported? (For analyses of each: 1 = yes, 0 = no). Differences were checked by a third researcher. No further adjudication was required.

Procedure

[7] www.qualtrics.com

The tests were administered in the order above. The CFMT+ was completed on a 15" laptop. The SFCT and Change Blindness videos were shown on a monitor (LG 32"LCD). Data for the *NASA-TLX Load Index*, *IPIP-NEO Personality Inventory* and Change Blindness questionnaires were collected on a PC, which also depicted the Change Blindness line-ups.

Results

Using IBM SPSS,[8] unless otherwise reported, alpha was maintained at $p = 0.05$, with post-hoc pairwise comparisons conducted using Tukey's. Very occasional missing data, were treated as missing on specific analyses only. Table 1 displays the correlation coefficients between all measures. To protect against Type-I errors from running multiple correlations, alpha here was set at $p = 0.01$. An explanation is provided in the relevant section below.

Table 1 about here

CFMT+: Table 2 displays mean scores by group, as well as numbers achieving the two SR thresholds. In total, 41 (16.3%) participants met the low SR threshold (90/102), 14 (5.6%), the high threshold (95/102). These high performances meant that the overall CFMT+ mean ($M = 75.2$, $SD = 12.8$) was significantly higher than in a recent representative UK study used to define the 95/102 SR threshold ($n = 254$, $M = 70.7$, $SD = 12.3$) (Bobak, Pampoulov, & Bate, 2016), one-sample test: $t(251) = 5.58$, $p < .001$. Nevertheless, there was no evidence of significant skewness (skewness = -0.23, SEM = 0.15, $p > .1$), and all participants were included in all correlational analyses (see Figure 2 for a CFMT+ score histogram).

[8] IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

A one-way ANOVA comparing mean CFMT+ scores by group (police experts, police identifiers, familiarised controls, unfamiliarised controls) was significant, $F(3, 247) = 21.29$, $p < .001$, $\eta^2 = .205$. Police experts outperformed police identifiers, although paired comparisons found only a marginal significant difference ($p = .092$). Both police groups outperformed familiarised and untrained controls ($p < .05$), who did not differ ($p > .05$).

Table 2 about here

Spotting a Face in a Crowd Test (SFCT): Table 3 displays the running time of each of the 11 video clips, the time on screen of each actor, mean hit rates, the approximate number of bystanders, as well as false positive (FP) identifications of bystanders, and the mean CR rate for clips when no allocated actors were depicted.

Table 3 about here

SFCT analyses were conducted in two components, as police experts were only allocated eight actors. Component 1 mainly employed two-way ANOVAs, as a function of group (police identifiers, familiarised controls, untrained controls) and actor-number (two, four, eight). Component 2, with fewer participants and lower statistical power, examined outcomes for participants provided with eight images only (police experts, police identifiers, familiarised controls, untrained controls). Table 4 displays SFCT scores by condition.

Table 4 about here

Completion Time:[9] The Component 1 ANOVA revealed significant group, $F(2, 221) = 32.90$, $p < .001$, $\eta^2 = .229$, and actor-number main effects, $F(2, 221) = 6.68$, $p = .002$, $\eta^2 = .057$, but no interaction, $F(4, 221) < 1$. Pairwise comparisons revealed police identifiers took significantly longer than the two control groups ($p < .05$). The control groups did not differ ($p > .2$). The eight-actor group took longer than four- and two-actor groups ($p < .05$), who did not differ ($p > .2$).

The Component 2 ANOVA was also significant, $F(3, 81) = 6.00$, $p = .001$, $\eta^2 = .182$. Both police groups took slightly longer than both control groups ($p < .06$ all comparisons).

Hits: The Component 1 ANOVA revealed significant group, $F(2, 235) = 9.89$, $p < .001$, $\eta^2 = .078$, and actor-number effects, $F(2, 235) = 18.20$, $p < .001$, $\eta^2 = .134$, but no interaction, $F(4, 235) = 1.60$, $p > .1$, $\eta^2 = .027$. Police identifiers' hit rates were higher than untrained controls ($p < .001$), and marginally higher than familiarised controls ($p = .058$). Two-actor group hit rates were higher than four- and eight-actor groups ($p < .001$), who did not differ ($p > .2$).

A significant Component 2 ANOVA, $F(3, 85) = 4.72$, $p = .004$, $\eta^2 = .143$ revealed higher police expert ($p = .008$) and identifier ($p = .021$) hit rates than untrained controls only.

False positives (FPs): The Component 1 ANOVA revealed a significant group effect, $F(2, 235) = 3.22$, $p = .042$, $\eta^2 = .027$. There was no actor-number, $F(2, 235) < 1$, or interaction effects, $F(4, 235) = 1.11$, $p > .2$, $\eta^2 = .019$. Familiarised controls made slightly fewer FPs than untrained controls ($p = .087$).

The Component 2 ANOVA was significant, $F(3, 85) = 3.25$, $p = .026$, $\eta^2 = .103$. Familiarised controls made fewer FPs than untrained controls ($p = .043$). Police experts made marginally fewer FPs than untrained controls ($p = .082$).

[9] Completion time data were not recorded for 14 participants.

CRs: The Component 1 ANOVA revealed significant group, $F(2, 235) = 3.07$, $p = .048$, $\eta^2 = .025$, and actor-number effects, $F(2, 235) = 6.22$, $p = .002$, $\eta^2 = .050$, but no interaction, $F(4, 235) = 1.78$, $p > .1$, $\eta^2 = .029$. Familiarised controls made slightly more CRs of empty clips than untrained controls ($p = .086$). CR rates were higher in the eight- than the four- and two-actor groups ($p < .05$ both comparisons).

The Component 2 ANOVA was not significant, $F(3, 85) = 2.05$, $p > .1$, $\eta^2 = .067$.

Confidence in Hits and FPs (Max = 3): A 2 (response type: hits, FP) x 3 (group) x 3 (actor-number) Component 1 mixed ANOVA on confidence ratings revealed a significant response type effect, $F(1, 216) = 187.30$, $p < .001$, $\eta^2 = .464$, no group, $F(2, 216) = 1.75$, $p > .1$, $\eta^2 = .016$; or actor-number effects, $F(2, 206) < 1$, but a significant response type x group interaction, $F(2, 216) = 20.13$, $p < .001$, $\eta^2 = .157$. Confidence in hits was higher ($M = 2.50$, $SD = 0.04$) than in FPs ($M = 2.03$, $SD = 0.04$). Simple effects analyses, $F(2, 239) = 18.44$, $p < .001$, $\eta^2 = .133$, found police identifiers had higher hits confidence than familiarised controls and untrained controls respectively (all comparisons $p < .05$). Police identifiers had lower confidence in FPs than untrained controls, $F(2, 223) = 3.46$, $p < .05$, $\eta^2 = .030$.

A 2 (response type) x 4 (group) Component 2 mixed ANOVA revealed a significant response type main effect, $F(1, 81) = 141.44$, $p < .001$, $\eta^2 = .636$; no group effect, $F(3, 81) = 1.67$, $p > .1$, $\eta^2 = .058$; and a significant interaction, $F(3, 81) = 9.84$, $p < .001$, $\eta^2 = .267$. Confidence was higher to hits ($M = 2.55$, $SD = 0.37$) than FPs ($M = 1.99$, $SD = 0.38$). Simple effects analyses of hits by group, $F(3, 85) = 5.75$, $p < .01$, $\eta^2 = .169$, revealed that police experts had higher confidence in hits than all other groups. Police identifiers had higher confidence in hits than untrained controls only (all comparisons $p < .05$). In contrast, simple effects on FPs by group, $F(3, 81) = 4.97$, $p < .01$, $\eta^2 = .155$, revealed police experts had lower confidence in FPs than all other groups ($p < .05$).

Correlations between CFMT and SFCT measures: From Table 1a, Pearson's tests revealed significant positive correlations between CFMT+ scores and SFCT completion time, hit rates, hits confidence, and CR rates. SFCT completion time also positively correlated with hit rates and hits confidence. In addition, there was a significant positive correlation between hit rates and confidence in hits, but no relationship between FP rates and FP confidence.

Table 5 about here

NASA-TLX Load Index: Table 5 displays the mean perceived SFCT workload outcomes by group. To measure Component 1, a 3 (group) x 3 (actor-number) MANOVA was conducted on the dependent variables of mental demand, temporal demand, effort, frustration level, and overall performance. The group effect was significant, Wilk's Lambda: $\Lambda = 0.910$, $F(10, 462) = 2.24$, $p = .015$, $\eta^2 = .046$. The actor-number, Wilk's Lambda: $\Lambda = 0.938$, $F(10, 462) = 1.49$, $p > .1$, $\eta^2 = .031$, and interaction effects, Wilk's Lambda: $\Lambda = 0.900$, $F(20, 767.1) = 1.23$, $p > .1$, $\eta^2 = .026$ were not.

Planned ANOVAs revealed that whereas the other four scale outcomes were not significant ($p > .1$), Overall Performance or belief in self-achievement was significant, $F(2, 235) = 9.57$, $p < .001$, $\eta^2 = 075$. Police identifiers provided the highest ratings on this scale in comparison to familiarised (marginally) ($p = .073$), and untrained controls ($p < .01$).

A Component 2 MANOVA by groups on the five dependent variables for Component 2 was not significant, $\Lambda = 0.795$, $F(15, 224.01) = 1.29$, $p > .2$, $\eta^2 = .073$.

From Table 1a, there were significant positive correlations between CFMT+ scores and the Temporal Demand, and Overall Performance measures, and negative correlations with Effort and Frustration Level. There were also significant positive correlations between Overall Performance and SFCT completion time, and confidence in hits. Participants taking

longer to complete the test had a stronger sense of achievement, and made their correct identifications, but not FPs, with higher confidence.

Table 6 about here

IPIP-NEO PI-R^{TM}: Table 6 displays the mean scores on each personality factor (extroversion, agreeableness, neuroticism, conscientiousness, openness) by group, as well as the published norms (Johnson, 2014). Five one-sample t-tests comparing norms to the combined control mean (familiarised, untrained) revealed only that current controls were more agreeable, $t(148) = 3.64$, $p < .001$; and conscientious, $t(148) = 2.25$, $p = .026$.

Five one-way ANOVAs examined the between-group data. There were no extroversion, $F(3, 243) = 1.96$, $p > .1$, $\eta^2 = 024$; or agreeableness effects, $F(3, 243) = 2.49$, $p > .05$, $\eta^2 = 030$. Those on neuroticism, $F(3, 243) = 6.49$, $p < .001$, $\eta^2 = 074$; openness, F(3, 243) = 8.13, $p < .001$, $\eta^2 = 093$; and conscientiousness, $F(3, 243) = 2.80$, $p = .040$, $\eta^2 = 034$ were significant. Post hoc analyses revealed no significant differences between police experts and any other group; or between the two control groups (all comparisons $p > .1$). Police identifiers scored significantly lower on neuroticism and openness than controls; and higher on conscientiousness than familiarised controls only ($p < .05$).

Table 1b demonstrates that in contrast to expectations there were no significant correlations between CFMT+ scores and any personality measure ($p > .1$),[10] and the only significant but weak negative correlations involving SFCT measures were between confidence in hits and openness, and SFCT completion time and neuroticism.

[10] As police recruitment strategies may target individuals possessing different personality characteristics to the general population (Twersky-Glasner, 2005), further analyses measured the relationship between the five personality measures and CFMT+ scores for controls only. Only the correlation between CFMT+ and openness (weakly) was significant, $r(149) = 0.18$, $p = .030$. The predicted correlation between extroversion and CFMT+ was not significant ($p > .05$).

Table 7 about here

Change Blindness: Responses on the two open-ended change blindness detection questions were coded as 1: change noticed, or 0: not noticed, and by group, Table 7 reports the proportion reporting the driver, clothing, both or no changes. Also included are event memory scores and line-up accuracy and confidence. All current participant groups exceeded the performances reported by Smart et al. (2014), with 79.4% here reporting the driver change (vs. Smart et al.: 46.5%), 40.7% the clothing change (vs. 17.8%), 35.9% both changes (vs. 9.9%), and 15.7% no changes (vs. 45.5%). Explanations are discussed below.

Four 2 (Reported: Yes, No) x 4 (group) chi-squared tests on the change blindness detection measures were significant, driver; $\chi^2(3, 248) = 9.36$, $p = .025$, Cramer's V = .194; clothes; $\chi^2(3, 248) = 15.67$, $p = .001$, Cramer's V = .251; both changes, $\chi^2(3, 248) = 18.46$, $p < .001$, Cramer's V = .273; or no changes, $\chi^2(3, 248) = 8.24$, $p = .041$, Cramer's V = .182. Post-hoc Fisher's Exact tests revealed no outcome differences between police groups, or between control groups (all comparisons $p > .2$). All police experts reported the driver change, although due to low power, police expert outcomes did not significantly vary from other groups ($p > .1$). Compared to control groups, police identifiers significantly more often reported clothing and both changes, while making more driver change reports, and fewer no change reports than untrained controls (all comparisons $p < .05$).

An independent measures ANOVA on event memory by group was not significant, $F(3, 244) = 1.68$, $p > .1$, $\eta^2 = .020$. From Table 1b, it can be seen that event memory scores positively correlated with reporting the driver change, clothes change, both changes, and negatively with no changes, suggesting that susceptibility to change blindness was mediated by attention to and memory of the video.

The four line-ups were scored as 1 = correct driver identification (target-present) or correct rejection (target-absent) and 0 = incorrect decision. Post-test, some participants reported being unsure as to whether they were supposed to identify Driver 1, Driver 2, or both. Therefore, as well as accuracy and confidence, participants were scored as to whether they made *at least* one correct target-present identification. There were no between-group target-absent line-up differences. Line-up 1, $\chi^2(3, 248) = 3.81$, $p > .2$, Cramer's V = .124. Line-up 4, $\chi^2(3, 248) = 3.82$, $p > .2$, Cramer's V = .124; or confidence in the first three line-ups, all $Fs \leq 1.68$, $p > .1$, $\eta^2 \leq .020$. However, the group effect was significant for the two target-present line-ups, Line-up 2, $\chi^2(3, 248) = 13.76$, $p = .003$, Cramer's V = .236; Line-up 3, $\chi^2(3, 248) = 13.50$, $p = .004$, Cramer's V = .233; and confidence in Line-up 4 (target-absent), $F(3, 240) = 2.91$, $p = .035$, $\eta^2 = .035$. However, the strongest effects were found on the analysis examining the likelihood of making at least one correct identification, $\chi^2(3, 248) = 25.17$, $p < .001$, Cramer's V = .319.

With Line-up 2, police experts were more likely than controls ($p < .05$), and marginally more likely than police identifiers to correctly identify Driver 1 ($p < .1$). With Line-up 3, untrained controls were less likely than police identifiers ($p < .05$), and marginally less likely than police experts to identify Driver 2 ($p < .1$). Both police groups were more likely to make at least one correct identification than untrained controls ($p < .05$ both comparisons) and marginally more than familiarised controls ($p < .1$ both comparisons). Surprisingly, the familiarised controls were marginally more likely to make at least one correct identification than the untrained controls ($p < .1$). With Line-up 4, police identifiers' confidence was higher than familiarised controls only ($p < .05$).

From Table 1b, there were significant positive relationships between CFMT+ scores and driver change reports and target-present and -absent line-up accuracy, and a negative relationship with no change reports, but no relationship with line-up confidence. There were

also positive correlations between SFCT hits confidence and clothing change detection, and the detection of both changes; as well as target-present and -absent line-up confidence. However, few other variables correlate suggesting the SFCT and Change Blindness tests may partly assess different skills.

Table 8 about here

Individual analyses: Table 8 reports the scores of the 14 participants achieving high CFMT+ SR criteria (95/102) on the primary dependent variables.[11] Also displayed are the mean scores of a subset of police identifiers, who scored *below* the low CFMT+ SR threshold (90/102), as previous research has assigned 90-94 scorers to SR groups. For the change blindness test, the proportions providing a 'yes' response to the detection measures, as well as making at least one target-present line-up identification of a driver are listed.

Where feasible, modified single case t-tests (Crawford et al., 2010), compared SR's performances against the police identifier subset mean.[12] Table 8 lists scores estimated to be above 98%, 95% and 90% of the population (of police identifiers). On the CFMT+ only, SRs scored higher than an estimated 95% of the population. On most other measures, the scores of only a few SRs differ substantially from the police identifier subset mean.

Discussion

This research found superior face recognition ability (*Cambridge Face Memory Test: Extended* (CFMT+); Russell et al., 2009), CCTV review experience, and target-actor

[11] Note: Only the SFCT scores of participants allocated eight target-actors in the SFCT scores are included, as task difficulty varied for those in the randomly allocated two- and four-actor conditions.
[12] No comparisons could be conducted on the five dichotomous change blindness variables.

familiarisation training enhanced performance at identifying actors in a bespoke videoed *Spot the Face in a Crowd Test* (SFCT), while also reducing *Change Blindness* susceptibility (Smart et al., 2014). Both tests partly replicated police CCTV footage reviews, while the SFCT also has similarities to searching for persons of interest at live events. However, motivations and procedures of searching for say terrorists, potential witnesses, or suspects will differ from one another, and from searching for actors described as missing persons in this study. As predicted however, on most measures, as a group, *police experts* who work in a full-time 'Super-Recogniser (SR) Unit' in London, outperformed other groups on the CFMT+, the SFCT and change detection measures, although, due to low police expert numbers, statistical significance was not always met. Front-line *police identifiers* also described as SRs by the MPS, outperformed controls on the CFMT+, and made more SFCT actor hits, and fewer change blindness errors. Individual analyses also compared the scores of all participants achieving high SR threshold (CFMT+ = 95+) (see Bobak, Pampoulov, & Bate, 2016), against a subset of police identifiers not meeting SR criteria (CFMT+ > 90). Apart from the CFMT+, on most measures, SRs did not significantly outperform this subset, primarily because between-groups analyses had greater power to detect outcome effects.

For all groups, correct SFCT actor identifications were associated with higher confidence than false positives (FPs) of bystanders, replicating the confidence-accuracy relationship in eyewitness research (for a review see Sauer & Brewin, 2015), although participants here simultaneously matched actor photos against footage. The superior hit rates of police were additionally associated with higher confidence than controls, while the FPs made by police, particularly experts, were made with lower confidence. This may reflect experience of searching indifferent quality CCTV footage for persons of interest. If no identification is made, a case may be closed and victims may not have access to justice. In contrast, a candidly tentative identification may provide the first lead to a missing person's

whereabouts, or in the case of a suspect, inculpatory or exculpatory evidence. These results suggest that protocols should allow for 'unsure' identifications at early investigative stages, without prejudicing future cases involving that police witness. An occasional incorrect tentative identification should not provide grounds to question any officer's reliability in the same, or a different case. Each identification should be assessed on its own merits.

The SFCT consists of eleven video clips, depicting two, one or zero actors. Except for police experts, always allocated eight actors, participants searched for two, four, or eight actors. Correct actor hit rates varied from 88% (Actor b in Clip D) to 9% (Actor g in Clip I), and CR rates of empty clips from 85% (Clip C) to 45% (Clip B) (see Table 2). Both police groups took longer to complete the SFCT than controls, suggesting greater meticulousness, which may partly explain their superior performance. Consistent with previous research (Tickner & Poulton, 1975), there was also a negative correlation between actor numbers and hit rates, although there was also a positive correlation between actor numbers and CR rates. Lower actor numbers required more 'clip empty' decisions, and successive 'empty' clips might have encouraged more guessing, meaning conclusions as to varying actor numbers are limited. However, there were no group or actor-number interactions. Police groups outperformed controls on the SFCT, regardless of numbers to search for.

To replicate real missing person searches, participants were provided with four photos of each actor, self-selected as typical of those received from families. The photos, up to a year old, depicted a variety of poses including full body shots. Clothing mostly differed in the SFCT videos, although some actors inadvertently occasionally wore matched or highly similar items in a photo (this was not planned). Indeed, Actor b wore a scarf in one photo and in clip D, associated with the highest accuracy (Table 3). Compared to single photos, viewing multiple images improves simultaneous face matching, probably due to better extrapolation of identity cues (Bindemann & Sandford, 2011; Dowsett Sandford, & Burton, 2015).

Furthermore, whole body images, can facilitate identification with unclear faces (Rice, Philips, Naro, An, & O'Toole, 2013). Nevertheless, performance reduces with appearance changes (Patterson & Baddeley, 1977), particularly with longer intervals between image acquisition (Megreya, Sandford, & Burton, 2013). However, many police reported the images replicated typical investigations, suggesting good test validity.

It should be noted however, that the individual analyses demonstrates that the SFCT possessed low power to discriminate between different ability performers, a consequence of the small number of critical outcome measures (e.g., maximum of 11 hits, and 2 CRs with eight target-actors) (See Table 8). An updated SFCT is required if used for deployment purposes. This might include additional video clips, and target-actors, as well as a visual narrative based on typical police work. Responses could be graded on detection of details, central and peripheral to this narrative.

Familiarisation training

Familiarised controls, in pairs, engaged in a pre-SFCT exercise by discussing actors' perceived personalities. Compared to untrained controls, this reduced FPs, and increased CR rates, but had no impact on hit rates, although confidence in hits was increased. This might be due to deeper social processing drawing attention to cues that assisted in bystander rejection (see also Bruce et al., 2001). However, other factors (e.g., motivation) from paired preparation rather than separately (untrained controls) might have enhanced performance, which have nothing to do with face processing. Indeed, these effects surprisingly carried over to the change blindness line-ups, prior to which there was no familiarisation exercise. Compared to untrained controls, there was a trend for familiarised controls to make at least one correct driver identification. Nevertheless, effect sizes were small and further research

should pinpoint this familiarisation effect locus. Indeed, there may be an upper limit, as familiarised controls did not outperform police on any SFCT measure.

Change Blindness

Police were also less susceptible to change blindness (Smart et al., 2014), in that more police (88.0%) than controls (73.8%) detected the switch of drivers, and of driver t-shirt colour (55.6% vs. 30.9% respectively) in the 2 min 44 sec video. Police also made more correct driver(s) identifications from subsequent line-ups, although there were no target-absent line-up differences. These effects were underpinned by face recognition ability, as there was a positive correlation between CFMT+ scores, driver change detection, and correct line-up identifications. However, far fewer participants in the original Smart et al. study reported the driver (46.5%) and clothing changes (17.8%) than even the control participants in the current study (over 70% and 30% respectively). One explanation is that despite spending an average of 45 min on the SFCT, instead of fatigue, current participants became used to analysing video, enhancing ability to spot anomalies. This suggests that the brief videos used in most change detection research may not capture police experience of multiple camera feed CCTV review operations that can take days. In the Smart et al. study there were also no differences between police and students, probably because unlike here, their police were not drawn from a select sample. These results suggest fears that police reviewers might be susceptible to change blindness if persons of interest move in and out of footage may be unfounded. Nevertheless, further research is required to investigate these effects using alternative paradigms, and to develop police protocols to reduce risks.

Perceived workload and personality

The perceived workload of the SFCT was measured using the *NASA-TLX Load Index* (Hart & Staveland, 1988). There were no between-group effects on mental demand, temporal demand, effort, or frustration level scales. However, compared to controls, police provided significantly higher overall performance ratings, indicative of higher self-achievement beliefs. Overall performance also correlated with SFCT completion time, and negatively with confidence in hits; which also independently negatively correlated with temporal demand. Participants slower to finish the SFCT displayed greater performance satisfaction. In contrast, those quicker to finish were less confident in SFCT identifications. The main contributors to negative workload perceptions are frustration, and mental and temporal demands (e.g. Warm et al., 1996). However, here, effect sizes were weak, and future research could investigate whether alternative measures may have uncovered reliable effects that might facilitate deployment decisions for such roles.

Participants also completed the *IPIP-NEO Personality Inventory* (Goldberg, 1999). Contrasting with previous research, there were no relationships between extraversion and face memory (e.g., Lander & Poyarekar, 2015); or between extraversion (e.g. Davies & Parasuraman, 1982) and neuroticism (e.g. Derakshan & Eysenck, 2009), with SFCT vigilance or visual search. The only significant correlations were between controls' CFMT+ scores and openness, and across all participants, SFCT hits confidence and openness. It is unclear why effects differed from past research, although police recruitment policy may value certain personality characteristics. Furthermore, the current controls were more conscientious and agreeable than the published norms. Despite this, police identifiers were more conscientious, and less neurotic and open than controls. Future research could test whether these measures could be used as selection criteria for deployment to police CCTV review operations.

Limitations

There are other limitations that might have had an impact. First, police tended to be white, male and over 30-years. In contrast, controls were mainly female, under 25, and more ethnically diverse. Face recognition ability peaks in the 30's (Germine, Duchaine, & Nakayama, 2011), and own-ethnicity faces are better recognised than other-ethnicity faces (Meissner & Brigham, 2001). Superior police performance here may reflect these differences. On the other hand, previous MPS police identifier research recruiting demographically matched controls revealed similar effects as here (Davis et al., 2016), and females tend to outperform males at face recognition (Bobak, Pampoulov, & Bate, 2016), particularly with female faces (Lewin & Herlitz, 2002). As the SFCT mainly contained female actors, the mainly female controls were from one of the UK's most ethnically diverse universities, likely to reduce cross-ethnicity effects; these factors were unlikely to strongly impact SFCT results.

Second, reviewing CCTV footage often involves more than searching for persons of interest. It may require plotting target movements across different camera feeds from different geographical locations, or identifying anomalous behaviours. By focussing on police with excellent face recognition skills only, and with no crimes depicted as in the SFCT or Change Blindness tests, there may be an adverse effect on operations requiring different skills – for instance, reviews of violent crimes or riots may be better served by those with relevant experience (see Damjanovic et al., 2014). Research should investigate a range of CCTV footage review requirements, to highlight which skills are associated with the best results.

Third, it might be assumed that police would be more motivated than controls, and therefore the results reported reflect motivational and not ability differences. However, anecdotal feedback revealed some police did not enjoy the challenge of these tests, and some blamed fatigue for what they considered to be poor performances – even though no

performance feedback was provided. Similar comments were elicited from police identifiers after surprisingly poor flowers recognition performances (Davis et al., 2016). Moreover, there is no evidence that controls were less motivated anyway. The current controls easily outperformed police in Smart et al.'s (2014) research on change blindness, and SFCT outcomes by the familiarised control group also suggest high engagement.

Selecting police for 'super-recognition' units and CCTV review roles

On the between-group analyses, police here mostly outperformed controls, supporting their inclusion in police identifier and expert groups. However, only a minority matched previous CFMT+ SR thresholds (32.3% achieved ≥ 90: Bobak, Bennetts et al., 2016: 11.1% achieved ≥ 95: Bobak, Pampoulov, & Bate, 2016). Some (*n* = 15; 14.2%) performed below the mean of one of the most representative UK samples to take the CFMT+ (Bobak, Pampoulov, & Bate, 2016). Yet, on other tests, many of the MPS pool have significantly outperformed controls (e.g. Robertson et al., 2016), and matched performances of high-threshold achieving SRs (Davis et al., 2016). Apart from the motivational factors and fatigue noted above, one reason may be that the design of the CFMT+, which mainly draws on short term memory of *unfamiliar* faces learned by their internal facial features only, may not match police identifiers' regular identifications of *familiar* suspects from full-body multiple stills and/or moving footage. Furthermore, the skills required by police experts who do regularly identify unfamiliar suspects may draw as much on their superior simultaneous face matching ability of faces in unconstrained images (see Robertson et al., 2016), rather than memory alone. Nevertheless, the results show that some police experts would definitely not be classified as SRs (CFMT+ ≤. 90). Even though no information as to how they were selected

to the unit is publically available, all were once members of the police identifier pool, and inclusion in that pool was mainly based on identification of multiple *familiar* suspects.

On the other hand, some high criteria CFMT+ (≥ 95) SRs score relatively poorly at other face processing tests such as simultaneous face matching (e.g., Bobak, Hancock, & Bate, 2016; Davis et al., 2016). This has been interpreted as evidence of potential SR sub-types (e.g. Bobak, Bennetts et al., 2016). However, the 72-trial original short version of the CFMT+ has been criticised for not measuring face-specific mechanisms, as some developmental prosopagnosics have recorded high scores (e.g., Esins, Schultz, Stemper, Kennerknecht, & Bulthoff, 2016; Horry, Cheong, & Brewer, 2014). As such, the 102-trial CFMT+ appears to provide a good marker for SR ability and for deploying police SRs to different roles. Nevertheless, further research could explore proposals made by Noyes et al. (2017) that a series of tests may be more suitable, in particular to rule out those with poorer face matching ability which may be an essential police skill in this context. These could include simultaneous face matching, long- and short-term face memory tests and a CCTV review and search task such as the SFCT.

Summary and conclusions

In summary, the results reported in this paper supports previous research finding that some MPS police possess superior face processing abilities (Davis et al., 2016; Robertson et al., 2016). Note however that these skills will not generalise to all police. Past research has mainly found no differences between police and the general public at eyewitness identification (e.g. Vredeveldt & van Koppen, 2016) or simultaneous face matching (e.g. Burton et al., 1999). The results also show that it might be risky for identifications from CCTV by police experts to be given higher weight in court than identifications by other

witnesses. Not all FPs by police in the current research were associated with low confidence, and MPS records show that police experts make occasional misidentifications. This is not surprising given the often poor quality of CCTV footage, and most errors are soon rectified after further investigation. However, all police may be susceptible to cognitive and confirmation biases which can encourage the interpretation of evidence conforming to pre-existing beliefs (see Edmund, Davis, & Valentine, 2015; Kassin, Dror, & Kukucka, 2013), and such risks may be enhanced with indifferent quality images.

London's MPS are the first force in the world to create a dedicated "SR unit", and their identification procedures follow the Police and Criminal Evidence Act (1984) Codes of Practice (Code D), designed to reduce miscarriages of justice. As with any forensic procedure, blind review of suspect identifications should be conducted to reduce risks. Legal researchers (e.g. Edmond & Wortley, 2016) have suggested that similar units could be created in other jurisdictions to ensure highest probity of CCTV identification evidence. The result of the current research would support those proposals, as police outperformed controls at face recognition, spotting faces in crowds and were less susceptible to change blindness. The outcomes have implications for any international police force intending to crate similar units, as there is a clear body of evidence demonstrating that police experts and identifiers can make disproportionate numbers of suspect identifications from CCTV. However, the protection of the rights of suspects must also be of paramount importance.

References

BBC News (2011). London riots: Most wanted suspect CCTV images released, downloaded 30 August 2016 from, http://www.bbc.co.uk/news/uk-england-london-16171972

Beck, M. R., Martin, B. A., Smitherman, E., & Gashen, L. (2013). Eyes-on training and radiological expertise: An examination of expertise development and its effects on visual working memory. *Human Factors and Ergonomics Society, 55,* 747-763. DOI: 10.1177/0018720812469224

Biggs, A. T., Cain, M. S., Clark, K., Darling, E. F., & Mitroff, S. R. (2013). Assessing visual search performance differences between transportation security administration officers and nonprofessional visual searchers. *Visual Cognition,* 21(3), 330-352. http://dx.doi.org/10.1080/13506285.2013.790329

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception, 40,* 625–627. http://dx.doi.org/10.1068/p7008

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

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

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5,* 339-360. http://dx.doi.org/10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*, 207–218. http://dx.doi.org/10.1037/1076-898X.7.3.207

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

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

Damos, D. L., & Bloem, K. A. (1985). Type A behaviour pattern, multiple-task performance, and subjective estimation of mental workload. *Bulletin of the Psychonomic Society, 23*, 53–56.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. London: Academic.

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

Davis, J. P., Lander, K., & Jansari, A. (2013). I never forget a face. *The Psychologist, 26*, 726-729. https://thepsychologist.bps.org.uk/volume-26/edition-10/i-never-forget-face

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.

Damjanovic, L., Pinkham, A. E., Clarke, P., & Phillips, J. (2014). Enhanced threat detection in experienced riot police officers: Cognitive evidence from the face-in-the-crowd

effect. *The Quarterly Journal of Experimental Psychology*, *67*(5), 1004-1018. http://dx.doi.org/10.1080/17470218.2013.839724

Derakshan, N., & Eysenck, M. W. (2009). Anxiety, processing efficiency, and cognitive performance: New developments from attentional control theory. *The European Psychologist, 14*, 168-176. http://dx.doi.org/10.1027/1016-9040.14.2.168

Donald F. M., & Donald, C. H. M. (2014). Task disengagement and implications for vigilance performance in CCTV surveillance. *Cognition, Technology and Work, 17*(1), 121-130. DOI: 10.1007/s10111-014-0309-8

Dowsett, A. J., Sandford, A., & Burton, A. M. (2015). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *Quarterly Journal of Experimental Psychology*, *69*(1), 1-10. http://dx.doi.org/10.1080/17470218.2015.1017513

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

Edmond, G., Davis, J. P., & Valentine, T. (2015). Expert analysis: Facial image comparison. In T. Valentine and J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 239-262). Chichester: Wiley-Blackwell.

Edmond, G., & Wortley, N. (in press). 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

Evison, M. P. (2014). The third forensics: Images and allusions. *Policing and Society: An International Journal of Research and Policy, 1,* 1-19. http://dx.doi.org/10.1080/10439463.2014.895347

Gelernter, J. (2013). Effective threat detection for surveillance. *IEEE International Conference on Technologies for Homeland Security*, Waltham, MA, pp. 290-296.

Germine, L., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118*(2), 201-210. http://dx.doi.org/10.1016/j.cognition.2010.11.002

Gibbs, R., Davies, G., & Chou, S. (2016). A systematic review on factors affecting the likelihood of change blindness. *Crime Psychology Review, 2*(1), 1-21. http://dx.doi.org/10.1080/23744006.2016.1228799

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lover-levels facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, and F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Hart, S. G., & Staveland, L. E. (1988). Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, and N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier.

Hillstrom, A., Hope, L., & Nee, C. (2008). Applying psychological science to the CCTV review process: a review of cognitive and ergonomic literature. London: HMSO.

Horry, R., Cheong, W., & Brewer, N. (2014). The Other-Race Effect in Perception and Recognition: Insights from the Complete Composite Task. *Journal of Experimental*

Psychology: Human Perception and Performance, *41*(2), 508-524. http://dx.doi.org/10.1037/xhp0000042

Ito, H., & Sakurai, A. (2014). Familiar and unfamiliar face recognition in a crowd. *Psychology, 5*, 1011-1018. http://dx.doi.org/10.4236/psych.2014.59113

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. Journal of Research in Personality, 51, 78-89. https://doi.org/10.1016/j.jrp.2014.05.003

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*, 42–52. doi:10.1016/j.jarmac.2013.01.001

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

Lander, K., & Poyarekar, S. (2015). Famous face recognition, face matching and extraversion. *Quarterly Journal of Experimental Psychology, 68*(9), 1769-1776. http://dx.doi.org/10.1080/17470218.2014.988737

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition — Women's faces make the difference. *Brain and Cognition, 50*(1), 121-128. http://dx.doi.org/10.1016/S0278-2626(02)00016-7

Li, J., Tian, M., Fang, H., Xu, M., Li, H., & Liu, J. (2010). Extroversion predicts individual differences in face recognition. *Communication and Integrative Biology, 3*, 295-298. http://dx.doi.org/10.4161/cib.3.4.12093

Manzoor, S. (2016). You look familiar: On patrol with the Met's super-recognisers. *The Guardian*, 5 November 2016, https://www.theguardian.com/uk-news/2016/nov/05/metropolitan-police-super-recognisers

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15*(5), 302-306. DOI: 10.1111/j.0956-7976.2004.00673.x

McGugin, R. W., Tanaka, J. W., Lebrecht, S., Tarr, M. J., & Gauthier, I. (2011). Race-specific perceptual discrimination improvement following short individuation training with faces. *Cognitive Science, 35*, 330–347. DOI: 10.1111/j.1551-6709.2010.01148.x

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology*, *25*, 30-37. http://dx.doi.org/10.1080/20445911.2012.739153

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology, 27*(6), 700-706. DOI: 10.1002/acp.2965

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*, 1-35. http://dx.doi.org/10.1037/1076-8971.7.1.3

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

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 406-417. http://dx.doi.org/10.1037/0278-7393.3.4.406

Police and Criminal Evidence Act (PACE) (1984). Codes of Practice (Code D). (2011). https://www.gov.uk/government/publications/pace-code-d-2011

Rice, A., Phillips, P. J., Natu, V., An, X, & O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science, 24*(11), 2235-2243. DOI: 10.1177/0956797613492986

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

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

Sadeh, N., & Bredemeier, K. (2011). Individual differences at high perceptual load: The relation between trait anxiety and selective attention. *Cognition and Emotion, 25,* 747-755. http://dx.doi.org/10.1080/02699931.2010.500566

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.

Sawin, D. A, & Scerbo, M. W. (1995). Effects of instruction type and boredom proneness in vigilance: implications for boredom and workload. *Human Factors, 37*(4), 752–765. dx.doi.org/10.1518%2F001872095778995616

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception, 28,* 1059-1074. doi:10.1068/p2952

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5,* 644-649. doi:10.3758/BF03208840

Singh, I. L., Tiwari, T., & Singh, A. L. (2007). Effects of target expectancy and cognitive demand on vigilance performance. *Journal of the Indian Academy of Applied Psychology, 33*(2), 151-156.

Smart, S. M., Berry, M. A., & Rodriguez, D. N. (2014). Skilled observation and change blindness: A comparison of law enforcement and student samples. *Applied Cognitive Psychology, 28,* 590-596. DOI: 10.1002/acp.3021

Temple, J. G., Warm, J. S., Dember, W. N., Jones, K. S., LaGrange, C. M., & Matthews, G. (1997). The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. *Human Factors, 42,* 183–194. DOI: 10.1518/001872000779656480

Tickner, A. H., & Poulton, E. C. (1975). Watching for people and actions. *Ergonomics, 18,* 35-51. http://dx.doi.org/10.1080/00140137508931438

Troscianko T., Holmes, A., Stillman, J., Mirmehdi, M., Wright, D. B., & Wilson A. (2004). What happens next? The predictability of natural behaviour viewed through CCTV cameras. *Perception 33*, 87-101. doi:10.1068/p3402

Twersky-Glasner, A. (2005). Police personality: What is it and why are they like that? *Journal of Police and Criminal Psychology, 20*(1), 56-67. DOI: 10.1007/BF02806707

Uenking, M. D. (2000). Pilot biofeedback training in the Cognitive Awareness Training Study (CATS). In *Proceedings of the AIAA Modeling and Simulation Technologies Conference*, Denver, CO, August 14-17, 2000.

Venkataramanan, M. (2015). The superpower police now use to tackle crime. *BBC Future.* Downloaded 4 November 2016 from http://www.bbc.com/future/story/20150611-the-superpower-police-now-use-to-tackle-crime

Vredeveldt, A., & Van Koppen, P. J. (2016). The thin blue line-up: A comparison of eyewitness performance by police and civilians. *Journal of Applied Research in Memory and Cognition, 5*, 252–256. doi: 10.1016/j.jarmac.2016.06.013

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 183–200). Mahwah, NJ: Erlbaum.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*, 433-441. DOI: 10.1518/001872008X312152

Wickens, C. D., & McCarley, J. D. (2008). *Applied attention theory*. CRC Press: London.

Wickens, C. D., Gempler, K., & Morphew, M. E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors, 2*(2), 99-126. http://dx.doi.org/10.1207/STHF0202_01

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*(7041), 439-440. doi: 10.1038/435439a

Figure 1: Images of each actor provided to participants prior to starting the SFCT (top row from left Actors A-D; bottom row E-H). Participants viewed these in advance and were able to directly match with images on the screen. Originally the four colour images of each actor were available on 210 × 297 mm paper. Note: one image is duplicated for Actors G and H.

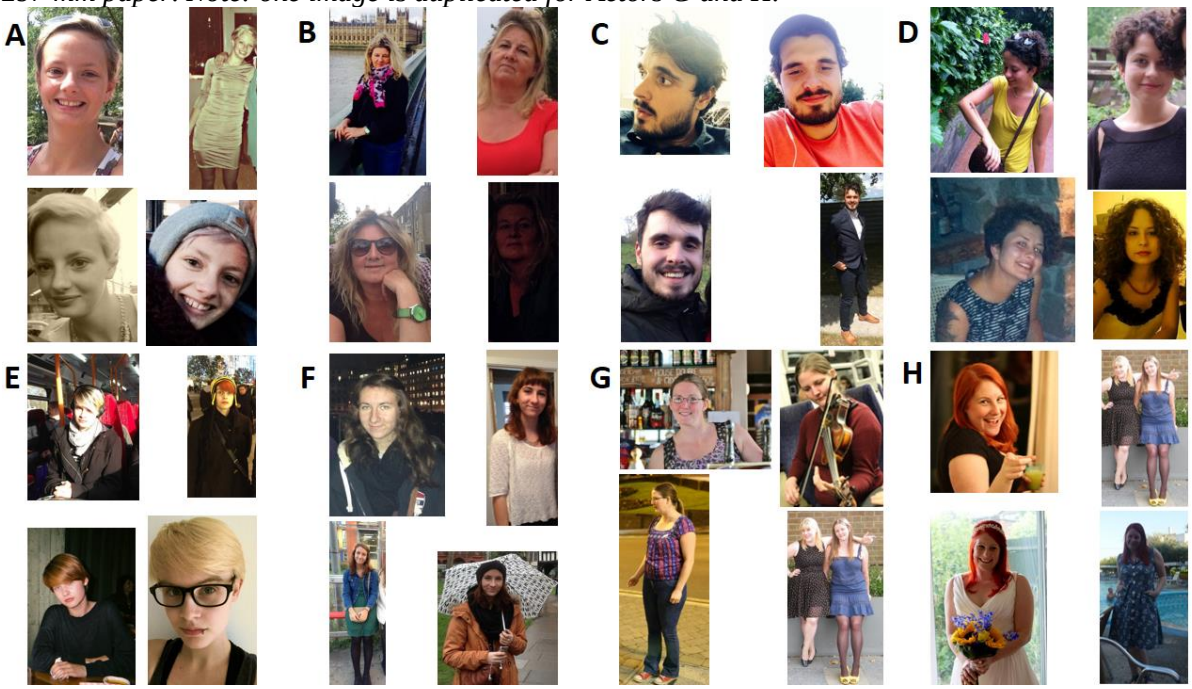

Figure 2: Histogram of CFMT+ scores

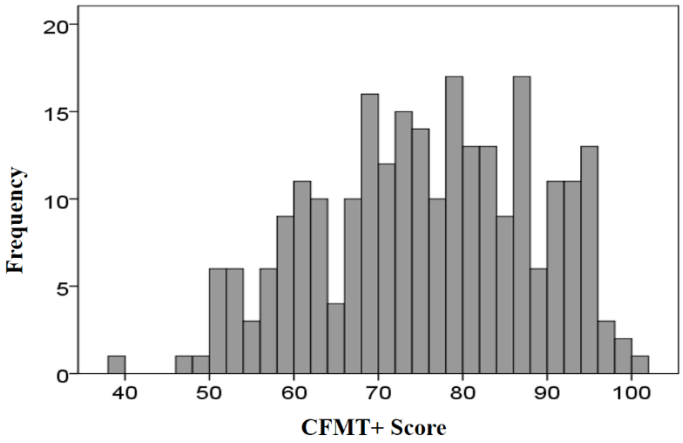

Table 2: Mean CFMT+ scores as a function of group (max = 102), with the numbers of participants meeting the two SR thresholds that have been employed in previous research

	Police Experts		Police Identifiers		Familiarised Controls		Untrained Controls	
n	7		92		62		90	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
CFMT+ score	90.4	8.0	81.5	10.7	72.2	11.4	69.8	12.5
	n	%	*n*	%	*n*	%	*n*	%
Score 95+ (high SR threshold)	3	42.9	8	8.7	2	3.2	1	1.1
Score 90+ (low SR threshold)	4	57.1	28	30.4	5	8.1	4	4.4

Table 3: Individual video clip (A-K), timing (min-sec), approximate bystander numbers (n), [A] target-actors (a-h), time of actor on screen in video clips (sec), hit rates (proportions), FP rates (proportions of participants making one or more bystander FPs in a clip) and CR rates (proportions by clip) to each actor and clip on the SFCT. See explanation in text.

	A		B	C	D	E	F	G	H	I	J	K	
Time (min)	1.47		1.58	0.54	1.11	1.34	1.32	1.48	2.12	1.46	1.34	2.04	
Byst-anders	10		12	14	16	28	19	16	6	32	12	17	
Actor	*b*	*f*	*d*	-	*b*	*h*	-	*f*	*c*	*g*	*e*	*a*	*d*
Screen	9	9	7	-	12	4	-	8	4	4	4	11	25
Hits	.82	.82	.86	-	.88	.24	-	.80	.87	.09	.64	.74	.83
FPs	.33		.69	.17	.24	.58	.25	.53	.25	.29	.35	.35	
CRs	.66		.45	.85	.67	.52	.77	.56	.67	.71	.59	.60	

[A] On most clips bystanders move rapidly across the screen and are replaced. In total, 100s passed across the screen over the course of the 11 clips. To calculate approximate bystander numbers, a count of all people depicted in the footage was taken by extracting a still at the mid-point of each clip.

Table 4: Performance outcomes on the SFCT as a function of group and image number

		Police experts		Police identifiers		Familiarised Controls		Untrained Controls	
n	Actor-Number	7		92		62		90	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
SFMT Completion Time (min)[A]	2	-	-	51.5	18.2	34.6	11.6	32.0	12.1
	4	-	-	56.5	23.6	37.0	14.5	33.0	9.9
	8	66.1	31.0	64.7	26.5	46.1	25.1	40.2	12.2
	Total	66.1	31.0	58.6	24.0	39.4	18.7	34.5	11.6
Proportion Hits	2	-	-	0.86	0.21	0.82	0.24	0.77	0.26
	4	-	-	0.76	0.23	0.55	0.24	0.56	0.19
	8	0.81	0.06	0.71	0.14	0.69	0.15	0.59	0.19
	Total	0.81	0.06	0.76	0.20	0.69	0.24	0.64	0.24
Hits Confidence (Max = 3)	2	-	-	2.72	0.35	2.43	0.43	2.37	0.40
	4	-	-	2.63	0.34	2.46	0.38	2.26	0.40
	8	2.86	0.08	2.62	0.32	2.57	0.38	2.32	0.38
	Total	2.86	0.08	2.65	0.33	2.49	0.40	2.32	0.39
Proportion Correct Rejections (CR)	2	-	-	0.58	0.29	0.71	0.27	0.66	0.22
	4	-	-	0.61	0.25	0.73	0.24	0.69	0.24
	8	0.86	0.24	0.82	0.27	0.86	0.23	0.68	0.29
	Total	0.86	0.24	0.69	0.29	0.77	0.25	0.68	0.24
False Positives (FP) (*n*) [B C]	2	-	-	4.50	3.27	3.15	3.17	3.75	2.59
	4	-	-	4.53	2.94	3.70	2.72	4.19	3.24
	8	2.29	1.80	4.11	2.49	3.00	1.88	5.41	4.54
	Total	2.29	1.80	4.35	2.83	3.27	2.59	4.33	3.43
False Positive (FP) Confidence	2	-	-	1.85	0.52	2.01	0.49	2.11	0.34
	4	-	-	1.99	0.53	1.94	0.32	2.18	0.36
	8	1.40	0.22	2.00	0.40	2.07	0.33	2.03	0.31
	Total	1.40	0.22	1.96	0.47	2.01	0.38	2.12	0.34

[A] Completion time data were not collected of some police identifiers (n = 4) and untrained controls (n = 10)
[B] With varying numbers of bystanders it was not possible to compute FP proportions
[C] Some participants made no false positives (FP) (n = 19)

Table 5: NASA-TLX Load Index outcomes as a function of participant group (image number data were collapsed as no effects were significant) [A]

	Police experts		Police identifiers		Familiarised Controls		Unfamiliarised Controls	
n	7		91		61		88	
	M	SD	M	SD	M	SD	M	SD
Mental Demand	2.43	0.98	2.73	1.09	2.95	1.08	2.87	1.14
Temporal Demand	2.29	1.48	1.67	1.20	1.59	1.25	1.57	1.16
Effort	1.57	1.13	2.03	1.22	2.32	1.05	2.39	1.01
Frustration level	0.43	0.79	0.91	1.27	0.76	1.05	0.99	1.20
Overall performance	3.43	0.71	2.62	1.30	2.15	1.31	1.75	1.32

[A] Note: Four participants provided no data on this test

Table 6: Mean scores as a function of participant group on the IPIP-NEO. Norms were also calculated based on 307,313 participants (Johnson, 2014, https://osf.io/tbmh5/) from the IPIP-NEO 300 dataset by extracting the first 30 answers to the short version used here [A]

	Police experts		Police identifiers		Familiarised Controls		Unfamiliarised Controls		Norms	
n	7		91		61		88			
	M	SD	M	SD	M	SD	M	SD	M	SD
Neuroticism	2.71	0.65	2.68	0.71	3.13	0.80	3.09	0.76	3.00	0.77
Extroversion	3.69	0.15	3.72	0.59	3.64	0.65	3.49	0.70	3.46	0.72
Openness	3.38	0.69	3.28	0.55	3.65	0.50	3.67	0.65	3.66	0.61
Agreeableness	3.14	0.56	3.58	0.52	3.71	0.53	3.58	0.61	3.46	0.57
Conscientiousness	3.76	0.43	3.89	0.43	3.64	0.56	3.71	0.64	3.57	0.64

[A] Note: Four participants provided no data on this test

Table 7: Change blindness, event memory and subsequent line-up outcomes as a function of participant group

	Police experts	Police identifiers	Familiarised Controls	Unfamiliarised Controls
n	7	92	61	88
	%	%	%	%
Driver Change	100.0	87.0	78.7	70.5
Clothing Change	42.9	56.5	32.8	29.5
Both Changes	42.9	52.2	29.5	22.7
No Changes	0.0	8.7	18.0	22.7
	M *SD*	*M* *SD*	*M* *SD*	*M* *SD*
Event Memory (out of 10)	6.32 1.45	6.54 1.43	6.36 1.33	6.08 1.51
	%	%	%	%
Line-up 1 (TA) Correct Rejections	85.7	52.2	60.7	53.4
Line-up 2 (TP) Correct IDs	57.1	22.8	13.1	10.2
Line-up 3 (TP) Correct IDs	71.4	51.1	39.3	27.3
Line-up 4 (TA) Correct Rejections	85.7	70.7	59.0	70.5
At least one correct ID (Line-ups 2 & 3)	85.7	67.4	49.2	33.0
	M *SD*	*M* *SD*	*M* *SD*	*M* *SD*
Line-up 1 Confidence	4.71 1.60	4.36 1.57	3.87 1.92	3.92 1.74
Line-up 2 Confidence	4.86 2.27	4.82 1.65	4.48 1.65	4.42 1.79
Line-up 3 Confidence	4.00 2.16	4.42 1.45	3.85 1.59	3.97 1.70
Line-up 4 Confidence	4.14 2.12	4.86 1.65	4.10 1.74	4.33 1.61

[A] Three participants failed to start the Change Blindness Test – one additional participant failed to provide data for Line-up 4 only

Table 8: Performances on each test of police experts (PE), police identifiers (PI), familiarised controls (FC) and untrained controls (UC) passing the higher CFMT+ SR threshold of 95/102, and mean performances of PI subset who scored below the lower SR threshold of 90/102 (SD in parentheses)

	Super-recognisers (SRs) scoring above 95 on CFMT+														PI
	SR1	SR2	SR3	SR4	SR5	SR6	SR7	SR8	SR9	SR10	SR11	SR12	SR13	SR14	
	PI	PE	UC	PE	PI	PI	PE	PI	PI	PI	PI	PI	FC	FC	Mean
Cambridge Face Memory Test: Extended (CFMT+)															*n* = 67
CFMT+	101***	99***	98***	97***	96**	96**	95**	95**	95**	95**	95**	95**	95**	95**	77.0 (9.30)
Spotting a Face in a Crowd Test (SFCT) [A]															*n* = 26
Hits (prop)	-	0.73	0.45	0.82	-	-	0.91*	-	-	-	0.82	0.64	0.73	0.82	0.73 (0.13)
CR (prop)	-	1.00	0.50	1.00	-	-	1.00	-	-	-	1.00	1.00	1.00	0.50	0.79 (0.29)
FP (*n*)	-	0*	3	4	-	-	4	-	-	-	2	2	1	3	3.92 (2.71)
Perceived Workload (NASA-TLX Load Index)															*n* = 67
Mental Demand	3.0	2.0	3.0	2.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	2.0	4.0	2.0	2.70 (1.07)
Temporal Demand	4.0***	1.3	2.7	1.3	2.7	1.3	2.7	1.3	0.0	4.0***	4.0***	1.3	2.7	2.7	1.55 (1.13)
Effort	1.0	3.0	1.0	3.0	2.0	3.0	1.0	2.0	3.0	2.0	1.0	0.0	1.0	3.0	2.10 (1.29)
Frustration Level	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	4.0***	0.0	0.0	0.91 (1.28)
Overall Performance	2.7	4.0	4.0	4.0	2.7	1.3	2.7	4.0	0.0	1.3	2.7	2.7	2.7	2.7	2.61 (1.32)
Personality (IPIP-NEO) [B]															*n* = 66
Neuroticism	2.3	3.3	3.5*	2.8	3.3	2.8	2.0	2.7	2.3	3.8**	3.5*	2.3	3.3	3.0	2.55 (0.70)
Extroversion	3.5	3.5	4.8**	3.8	4.0	4.2	3.5	3.7	3.0	3.7	3.8	3.3	3.3	3.7	3.74 (0.57)
Openness	3.8	2.5	4.8***	3.5	3.2	3.5	4.2*	3.7	2.0	4.3**	3.7	2.5	3.7	3.8	3.24 (0.58)
Agreeableness	3.7	2.8	4.0	3.8	3.7	3.5	3.5	4.0	2.2	4.0	2.8	3.3	4.2	4.2	3.57 (0.52)
Conscien-tiousness	4.2	3.8	4.2	3.8	3.8	4.0	4.5	4.0	3.5	3.7	4.2	4.5	3.8	3.8	3.90 (0.51)
Change blindness, event memory and line-up outcomes															*n* = 67
Driver [C]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	0.85 [D]
Clothes [C]	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No	Yes	No	No	0.55 [D]
Both driver and clothes [C]	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No	Yes	No	No	0.51 [D]
None [C]	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	0.10 [D]
Event Memory	9.0*	8.0	6.0	7.0	7.0	8.0	3.0	8.0	5.0	6.0	5.0	8.0	8.0	7.0	6.39 (1.46)
At least one TP ID [C]	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	0.63 [D]
TA Accuracy	1.00*	0.50	1.00*	1.00*	0.00	0.00	1.00*	0.50	1.00*	0.50	0.50	0.50	1.00*	1.00*	0.54 (0.40)

*** Estimated (*** > 98%; ** > 95%; * > 90%) of population (police identifier) falling below individual SR's score

[A] SFMT performances are only reported for SRs and PIs provided with eight target-actor images, meaning that some SRs (*n* = 6) and members of the PI subset (*n* = 41) provided with two or four target-actor images were excluded

[B] One PI provided no data on the IPIP-NEO

[C] Yes = detected driver, clothes, both, detected no changes, and/or made at least one target-present correct driver identification

[D] Proportion of PIs being scored with a 'Yes' response

Table 1a: Correlation coefficients between all outcome measures

		SFCT						Perceived Workload				
		Time	Hits	Hits Confidence	FA	FA Confidence	CR	Mental Demand	Temporal Demand	Effort	Frustration Level	Overall Performance
CFMT+		.30 *	.18 *	.27 *	-.13	-.12	.17 *	<.01	.19 *	-.19 *	-.19 *	.26 *
SFCT												
	Time		.22 *	.25 *	.13	.02	-.02	-.09	-.02	-.06	.03	.18 *
	Hits			.17 *	.02	-.03	-.06	.07	.11	-.12	-.05	.08
	Hits Confidence				-.03	.04	.01	<.01	.15	-.08	-.10	.22 *
	FP					-.03	-.73 *	-.08	.02	.02	.07	-.08
	FP Confidence						.03	.05	.02	.09	-.09	-.06
	CR							.04	.03	-.05	-.03	.03
Perceived workload												
	Mental Demand								-.05	-.15	-.17 *	-.29 *
	Temporal Demand									-.30 *	-.49 *	.24 *
	Effort										-.10	-.27 *
	Frustration Level											-.31 *

Note: To protect against Type-I errors associated with multiple tests, criteria for significance was $p < .01$ (marked with *)

Table 1b: Correlation coefficients between all outcome measures for all participants. Note: there were occasional missing data on analyses involving workload, personality and change blindness measures

	Personality					Change Blindness						Line-ups		
	Neuroticism	Extroversion	Openness	Agreeable-ness	Conscientious-ness	Driver	Clothes	Both	No Change	Event Memory	At least 1 TP Correct ID	TP Confidence (M)	TA Accuracy (M)	TA Confidence (M)
CFMT+	-.09	.06	.01	-.05	.02	.21 *	.13	.16	-.19 *	.13	.27 *	.12	.20 *	.15
SFCT														
Time	-.22 *	.07	-.09	-.05	.10	.05	.11	.12	-.04	.12	.22 *	.20 *	.06	.16
Hits	-.12	-.06	-.08	-.05	.02	.09	.16	.16	-.10	.06	.13	.06	.03	.03
Hits Confidence	-.10	.06	-.17 *	.02	.10	.13	.20 *	.20 *	-.14	.05	.15	.35 *	.13	.29 *
FP	-.11	-.05	-.10	-.06	<.01	-.14	-.05	-.13	.06	-.16	.02	-.07	-.04	-.01
FP Confidence	-.02	-.07	-.02	.08	.09	-.06	-.01	-.03	.04	.09	.03	.12	-.03	.15
CR	.03	-.09	.10	.10	-.03	.05	.01	.04	-.01	.17 *	-.02	.07	<.01	.05
Perceived workload														
Mental Demand	-.08	.04	.07	.06	.02	-.02	-.06	-.07	.02	.02	.04	-.02	.03	-.04
Temporal Demand	-.04	.02	-.02	.03	.01	.07	-.05	-.03	-.05	-.01	.04	<.01	.06	-.01
Effort	<.01	<.01	.12	<.01	-.01	-.10	-.16	-.17 *	.10	-.08	-.16	.01	-.11	.08
Frustration Level	.13	-.08	-.08	-.14	-.13	-.09	.08	.06	.07	-.07	.04	-.09	-.03	-.09
Overall performance	-.02	.05	-.10	.07	.13	.15	.14	.16	-.14	.11	.11	.11	.13	.11
Personality														
Neuroticism		-.13	.09	.08	-.07	-.07	-.10	-.10	.08	<.01	.19 *	-.21 *	-.06	-.19 *
Extroversion			.23 *	.07	.30 *	.03	-.01	-.01	-.03	-.06	.02	.11	.06	.16
Openness				.19 *	.12	-.01	.02	<.01	-.02	<.01	.09	-.04	-.07	-.05
Agreeableness					.35	-.03	.07	.02	-.04	.06	-.06	.02	-.03	.09
Conscientiousness						-.02	.04	.03	<.01	-.05	.06	.26 *	.09	.30 *
Change Blindness														
Driver							.18 *	.38 *	-.85 *	.18 *	.16	.10	.19 *	.19 *
Clothing								.90 *	-.36 *	.25 *	.17 *	.06	.13	.04
Both									-.32	.22 *	.18 *	.08	.18	.08
No change										-.26 *	-.18 *	-.09	.14	-.16
Event memory											.14	-.03	-.04	.02
Line-ups														
At least one target present (TP) correct identification												.08	.59 *	.13
TP Confidence													.15	.65 *
Target absent (TA) mean accuracy														.22 *

[A] Note: Data for the Change Blindness target-absent Line-up 1 and Line-up 4 tests have been combined in Table 1. The target-present accuracy outcomes are based on participants making at least one identification (see text).

Note: To protect against Type-I errors associated with multiple tests, criteria for significance was $p < .01$ (marked with *)