Received: 30 April 2018 | Revised: 10 January 2019 | Accepted: 15 January 2019

DOI: 10.1002/acp.3525

RESEARCH ARTICLE

WILEY

The consistency of superior face recognition skills in police officers

Sarah Bate[1] | Charlie Frowd[2] | Rachel Bennetts[3] | Nabil Hasshim[1] | Emma Portch[1] | Ebony Murray[1] | Gavin Dudfield[4]

[1] Department of Psychology, Bournemouth University, Poole, UK

[2] School of Psychology, University of Central Lancashire, Preston, UK

[3] College of Health and Life Sciences, Brunel University, London, UK

[4] Dorset Police, Bournemouth, UK

Correspondence
Sarah Bate, Department of Psychology, Faculty of Science and Technology, Poole House, Bournemouth University, Fern Barrow, Poole BH12 5BB, UK.
Email: sbate@bournemouth.ac.uk

Funding information
British Academy Mid-Career Fellowship, Grant/Award Number: MD170004

Summary

In recent years, there has been increasing interest in people with superior face recognition skills. Yet identification of these individuals has mostly relied on criterion performance on a single attempt at a single measure of face memory. The current investigation aimed to examine the consistency of superior face recognition skills in 30 police officers, both across tests that tap into the same process and between tests that tap into different components of face processing. Overall indices of performance across related measures were found to identify different superior performers to isolated test scores. Further, different top performers emerged for target-present versus target-absent indices, suggesting that signal detection measures are the most useful indicators of performance. Finally, a dissociation was observed between superior memory and matching performance. Super-recognizer screening programmes should therefore include overall indices summarizing multiple attempts at related tests, allowing for individuals to rank highly on different (and sometimes very specific) tasks.

KEYWORDS

composite face processing, face recognition, individual differences, personnel selection, super recognizers

1 | INTRODUCTION

In the last decade, there has been growing interest in so-called "super-recognizers" (SRs): people with an extraordinary ability to recognize faces (Bobak, Hancock, & Bate, 2016; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016; Russell, Duchaine, & Nakayama, 2009). Although much of the published work examining these individuals has theoretical intentions (e.g., Bate & Tree, 2017; Bennetts, Mole, & Bate, 2017; Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Bobak, Parris, Gregory, Bennetts, & Bate, 2017; Ramon, Miellet, Dzieciol, Konrad, & Caldara, 2016; Russell, Chatterjee, & Nakayama, 2012), there has been increased applied interest in the deployment of SRs in policing and security settings. Yet the published literature lacks any large-scale investigations into the consistency of superior face recognition skills either within or across tasks, with most studies merely requiring performance at an arbitrary level on a single task for inclusion in an SR sample (see Bate et al., 2018). It is therefore unknown whether individuals with genuine proficiencies are being detected: This not only draws existing theoretical work into potential disregard but also has implications for the performance of SRs in real-world settings.

The extended version of the Cambridge Face Memory Test (CFMT+; Russell et al., 2009) is currently the dominant test used to detect super recognition, and the sole inclusion criterion used in many papers is a single attempt at this test where the score exceeds control performance by at least two standard deviations (see Bobak, Pampoulov, & Bate, 2016). The protocol of using a single inclusion

criterion based on a somewhat arbitrary statistical cut-off is problematic. Although some individuals may simply reach criterion by chance, others, who are genuinely excellent at face recognition, may be "missed." The latter may occur because of fatigue, illness, lifestyle influences, or simple misunderstanding of instructions—factors that may be overcome by repeated assessment. A similar scenario has been noted at the other end of the face recognition spectrum, where McKone et al. (2011) carried out a second screening session to clarify the diagnoses of six individuals who reported severe everyday difficulties with face recognition. Although these people only achieved borderline impaired scores in an initial assessment, they did fulfil the criteria for prosopagnosia in a second attempt at the test using novel stimuli. In another study, Bindemann, Avetisyan, and Rakow (2012) examined performance consistency in typical participants who completed the same face matching task on three subsequent days. They found that individual participants varied in their overall accuracy scores on each day, eliciting different responses to the same stimuli across the three attempts. Thus, repeated assessment of performance on the same task may be required to (a) interpret borderline cases and (b) detect not only the most proficient but also the most reliable performers.

Much existing evidence also suggests that an individual's genuine level of performance may differ across face recognition tasks that tap into different subprocesses. For instance, some people may be very good at discriminating between simultaneously presented faces, yet only have average face memory skills. Evidence supporting this possibility comes from the developmental prosopagnosia literature, where dissociations between subcomponents of face recognition have been observed. Although impaired face memory is the hallmark symptom of the condition (Murray, Hills, Bennetts, & Bate, 2018), earlier processes involving the perception of faces can be selectively spared (Bate, Haslam, Jansari, & Hodgson, 2009; Lee, Duchaine, Wilson, & Nakayama, 2010; McKone et al., 2011) or impaired (Bate et al., 2009; Chatterjee & Nakayama, 2012; Duchaine, Germine, & Nakayama, 2007; for a review, see Bate & Bennetts, 2015). Interestingly, some small-scale investigations into super recognition have found that facial identity perception (typically assessed via face matching tasks that place no demands on memory) is not always facilitated in individuals with superior face memory skills (Bate et al., 2018; Bobak, Bennetts, et al., 2016; Bobak, Dowsett, & Bate, 2016), although it is unclear whether the reverse pattern can be found (i.e., facilitated face matching skills in the context of typical face memory skills). This is because performance on a face memory task (the CFMT+) is typically the sole screening measure for theoretical investigations, and face perception skills have only been reliably assessed in individuals who have passed the initial inclusion criterion.

Importantly, screening procedures that use the CFMT+ alone also ignore another fundamental indicator of face recognition performance: the ability to decide when a target face is absent from an array. Yet face recognition in the real world, and particularly within policing settings, does not only involve the recognition of a target face when it is *present* within a set of faces but also importantly also requires successful acknowledgement that a particular face is *absent*. Although top performers should demonstrate heightened performance in both scenarios, some existing evidence indicates variation in target-absent accuracy in SRs who had initially been identified by the CFMT+ (i.e., target-present performance) alone (Bobak, Hancock, et al., 2016). Given work with typical participants has also failed to find an association between target-present and target-absent face matching performance (McCaffery, Robertson, Young, & Burton, 2018; Megreya & Burton, 2007), inclusion of both measures within a screening test is necessary to provide a complete indicator of top-end face recognition performance.

Finally, most traditional face recognition tasks use tightly controlled facial images that have been stripped of external features that could cue recognition (e.g., Bate, Haslam, Tree, & Hodgson, 2008; Duchaine & Nakayama, 2006; McKone et al., 2011). However, some authors suggest that this adjustment reduces ecological validity by failing to replicate the immense variability that typically occurs between different images of the same face in everyday life (Young & Burton, 2017, 2018). In fact, the matching of two unfamiliar faces of the same identity is a notoriously difficult task (e.g., Jenkins, White, Van Montford, & Burton, 2011; Young & Burton, 2017, 2018), even when external features are present and the two images have been collected on the same day (e.g., Bruce et al., 1999). The task becomes even more difficult when images have been captured on different days, and in this instance, the inclusion of extra-facial features can serve to further increase variability between naturalistic images (e.g., where the target has changed hairstyle, grown facial hair, or is wearing alternative make-up). For example, Kramer and Ritchie (2016) examined the influence of glasses on face matching performance. They found that typical participants incorrectly categorized more same-identity pairs when glasses were worn in only one image, compared with pairs where they were worn in both or neither image. Embracing real-world variability in facial presentation may therefore not only be an important means of replicating real-world policing scenarios (particularly where individuals may deliberately attempt to disguise their identity) but may also enhance the difficulty of face recognition tasks, ensuring they are appropriately calibrated for the detection of top performers.

The current study aimed to examine the consistency of superior face recognition skills both across tests that tap into the same process and between tests that assess different processes. We assessed the performance of a group of 30 police officers who had previously been screened for super recognition, surpassing a liberal criterion on at least one of two tests: the CFMT+ and a face matching task. This allowed us to assess face recognition consistency in those with apparent proficiencies in both memory and matching, in addition to those with facilitations in only one of the two processes. All officers completed five tests: a new face memory test that adapted the CFMT+ paradigm to include target-absent trials (Bate et al., 2018), three new versions of the face matching task, and a test that requires participants to decide whether a composite target face (generated using a holistic composite system) is present within a simultaneously presented image displaying a crowd of people ("Crowds" task). We included the Crowds test to examine whether proficient face recognition skills, as identified on either of the two preceding types of test, extend to a novel, more real-world policing task. All tests were calibrated to detect performance at the top end of the spectrum (allowing for at least three standard deviations from the control mean), using naturalistic facial images that varied in appearance. Consistency of performance across related tests was considered in terms of the number of times that a participant surpassed criterion performance and by overall index scores.

2 | METHOD

2.1 | Participants

Thirty police officers (10 female, M_{age} = 37.6 years, *SD* = 7.9) from the United Kingdom took part in this study. These officers had previously been identified as having proficient face recognition skills following a large-scale screening programme carried out by our laboratory (see Data S1). Because we wanted to identify individuals who were proficient at face memory *or* face matching, these officers had obtained excellent scores on at least one of two tests: the CFMT+ (for full details, see Russell et al., 2009) and a face matching task (the Pairs Matching Test [PMT]; see Bate et al., 2018). Although the CFMT+ is a well-known test, the PMT is a more recent test developed within our laboratory. A detailed description of the latter test can be found in Bate et al. (2018); in brief, the PMT has a similar design to existing face matching tasks (e.g., Burton, White, & McNeil, 2010), but is sufficiently calibrated to detect top performers via single-case statistical comparisons. The task contains 48 (half male) pairs of faces, presented in colour. Half of the trials match in identity, and half are mismatched. Each pair of faces is displayed simultaneously for an unlimited duration, and participants elicit a "same" or "different" response for each pair.

Because each officer only had one attempt at each test, we set the selection cut-off at 1.5 *SD*s above the control mean (see Data S1). Although this liberal criterion is lower than that used in previous work, it allowed borderline cases to be included—enabling us to thoroughly examine the importance of repeated testing and performance consistency. Using these cut-offs, 14 officers outperformed controls on both tests, 10 only on the CFMT+, and six only on the PMT. Twenty-eight officers were Caucasian; two were of mixed ethnicity. These individuals perform a wide range of roles within the police force, with 21 having direct contact with the general public. Length of service ranged from 1 to 31 years. Officers participated in this investigation during their normal working hours and did not receive any additional compensation for their time.

Forty (20 female; *M* = 33.4 years, *SD* = 10.2) civilian control participants, age-matched to the police participants, also took part in this study. They were randomly selected from Bournemouth University's participant pool, irrespective of their self-perceived face recognition skills. These individuals were offered a small financial incentive to ensure their motivation for the tasks. Ethical approval for the study was granted by the institutional ethics committee.

2.2 | Materials

2.2.1 | Models Memory Test

This new test of face memory is an adaptation of the CFMT+, using naturalistic colour photographs of each individual that have been captured on different days and in different settings (see Figure 1). Images are cropped to display the faces from the neck upwards (image sizes are 8 cm high by 6 cm wide), but no external facial features are removed.

A full description of the Models Memory Test (MMT) can be found in Bate et al. (2018). In brief, the test begins with a similar encoding procedure to the CFMT+: For each of six target faces, three different images of the person (taken on different days and in different settings) are shown sequentially for 3 s and immediately followed by three test trials. Three faces are displayed in each test trial: one of the encoded images and two distractors. As in the CFMT+, the encoding phase terminates with a 20-s review of the six target faces, by simultaneously presenting a new frontal image of each individual.

Ninety test trials (45 target-present) are then presented in a random order, with a screen break at the halfway point. Target-present triads contain one new image of a target face and two matched distractors; target-absent triads contain three distractors that are matched to one of the target faces. Triads in the first half of the test contain images that more closely resemble those used in the encoding phase, whereas those presented after the screen break display the targets under more challenging conditions (e.g., with additional facial hair, or where the face was obscured by accessories or a large change in viewpoint).

Images remain on-screen until a response is made, and no time restriction is imposed. Participants can make a target-present or target-absent response for each trial. Target-present responses were elicited using the corresponding number key (1–3) that indicates the position of the target in the triad, whereas the 0 key represents a target-absent response. Five types of response are possible on this test. For target-present trials, participants can correctly identify the target face (hits), they can incorrectly elicit a target-absent response (misses), or they can incorrectly identify one of the distractor faces (misidentifications). In target-absent trials, participants can elicit the correct response (correct rejections) or incorrectly identify a distractor face (false positives). We recorded each of these responses for each participant and summed the number of hits and correct rejections to calculate an overall accuracy score.

2.2.2 | Pairs Matching Test

Three new blocks of the PMT (see Data S1 and Bate et al., 2018) were developed for this investigation. These assessed participants' ability to match simultaneously presented pairs of male Caucasian faces when (a) the viewpoint of the face severely changed (i.e., by more than 45°) across the two images, (b) the actor was wearing glasses in only one image, and (c) the actor had facial hair in one image but was cleanly shaven in the other (see Figure 2). Each of these three blocks contained 48 trials: 24 matched in identity, whereas the remainder displayed two different individuals. All images were downloaded from Google image searches and were cropped to display the entire face from the neck upwards. Mismatched faces were paired according to their perceived similarity to each other, and all images were adjusted to 10 cm in width and 14 cm in height. Participants completed the three blocks in a counterbalanced order, and trials were randomized within each block. To ensure ecological validity (i.e., in replicating policing scenarios such as CCTV image matching), stimuli were displayed until responses were made, and no time limit was imposed. Participants made key presses to elicit "same" or "different" responses. Scores were calculated in terms of hits (the number of correct "same"

FIGURE 1 Sample stimuli from the MMT. Note that these trials are all target-present. Due to issues with image permissions, this figure only displays images that resemble those used in the actual test

responses) and correct rejections (the number of correct "different" responses) and summed for overall accuracy.

2.2.3 | Crowds Matching Test

Our final test aimed to replicate a very specific policing scenario, where officers have a composite target face (generated using EvoFIT: a holistic composite system) and they are required to find this individual in a crowd. A detailed description of this test and the composite generation procedure can be found in Bate et al. (2018) and is also summarized in Supporting Information (see Data S2). In brief, an initial set of participants (see Data S2) generated the target composite stimuli, following a pre-existing procedure (Fodarella, Kuivaniemi-Smith, Gawrylowicz, & Frowd, 2015). This process began with participants freely describing a designated target face (half taken from the crowd images used in the final test and half taken from crowd images that were not used in the final test) in as much detail as possible, without guessing. This information was recorded by the experimenter on a face-description sheet, using feature description labels. An age- and gender-appropriate database was then presented to the participant, displaying the inner region of a series of faces. Participants selected faces that best matched the overall appearance of the target; these faces were combined, and the selection procedure repeated. They then selected the best-matching item and improved it using "holistic" (addressing the age, weight, and overall appearance of the face) and "shape" (addressing the size and position of facial features) tools. Finally, the best-matching set of external features (hair, ears, and neck) were selected, and participants had a final opportunity to improve the face using the same holistic and shape tools.

Thirty-two composites were selected for the final experiment (see Data S2) and encompassed into 32 trials (16 target-present) where participants simultaneously viewed a target composite face at the top of the screen and an image below that showed 25–40 people in a naturalistic setting (e.g., an audience at a concert or sporting event; see Figure 3). Composite faces measured 3 cm in height and 2 cm in width, and crowd images were 9 cm in height and 13 cm in width. Participants were required to decide whether or not the target face is present in each crowd, pressing a key on the keyboard to make their response. Trials were displayed in a random order, with no time restriction for responses. Hits and correct rejections were calculated and summed for overall accuracy.

2.3 | Procedure

The majority of the officers was tested in face-to-face laboratory conditions. However, due to limitations in availability, a minority of individuals (N = 5) completed some or all of their testing

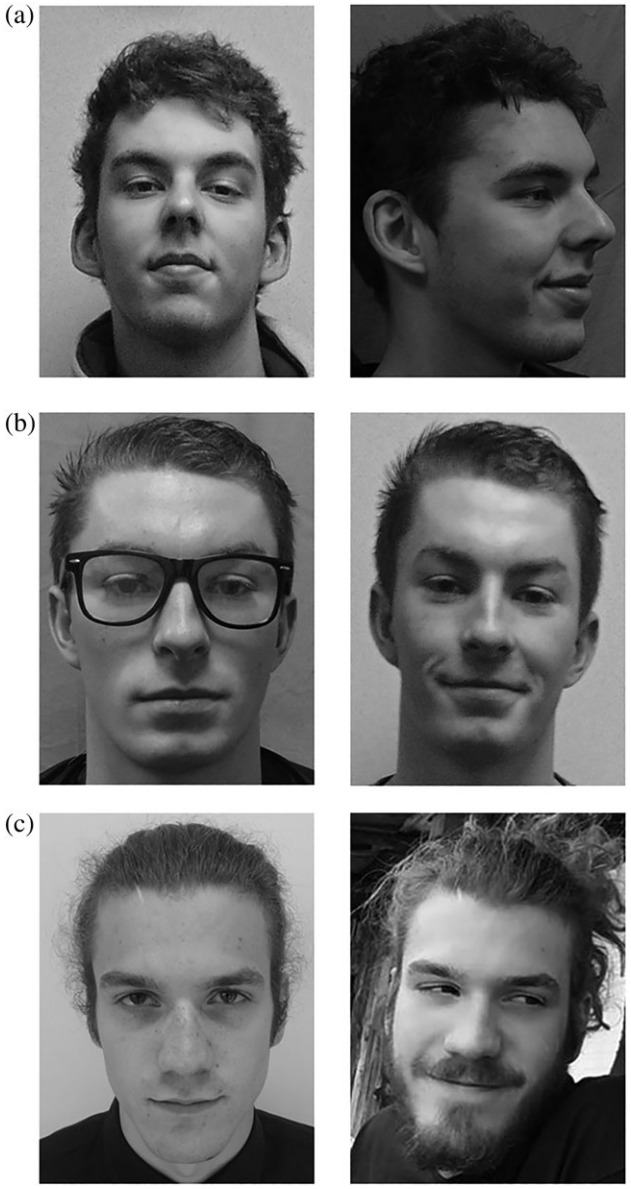

FIGURE 2 Sample pairs from the three new blocks of the PMT that differ according to (a) pose, (b) glasses, and (c) facial hair. Due to issues with image permissions, this figure only displays images that resemble those used in the actual test. All pairs display faces of the same identity

session online (via a testing platform on our laboratory's website: www.prosopagnosiaresearch.org). As tests were completed in a counterbalanced order, this affected different tests for different individuals. To allow for the possibility that performance may vary for online versus laboratory conditions, half of the control participants completed the tests remotely, and the remainder took part under strict experimental conditions.

2.4 | Statistical analyses

Initial analyses compared the performance of online versus laboratory control participants. As no differences were detected on any measure (all *ps* > 0.05), data were collapsed across all control participants for subsequent analyses. For all tests, the overall mean and SD scores were calculated for all performance measures, and cut-offs in this phase were set at the usual, more conservative level of 1.96 *SD*s from the control mean. Because all the tests contained target-present and target-absent trials, these items were also analysed separately, together with relevant signal detection measures (see below for each test). Initial exploration of the data revealed that one officer scored 97.78% correct on the target-present trials of the MMT, but made no correct responses on the target-absent trials. We assumed this individual had misunderstood the task and removed their data from all relevant analyses.

3 | RESULTS

3.1 | Relatedness of tests

The main aim of this investigation was to examine consistency of performance across tests that tap the same process and between tests that measure different processes. Initial analyses therefore collapsed data across SR and control participants and explored the relationship between the experimental tests and the CFMT+. Further, because existing work (e.g., Bate et al., 2018) has indicated differences in target-present and target-absent performance in super recognition, we entered data for each test separately for hits and correct rejections.

Initial eigenvalues from a principal components analysis (PCA) indicated that the first three factors explained 33.57%, 23.39%, and 10.71% of the variance, and the remaining eight factors had eigenvalues that were less than 1. Solutions for two, three, four, five, and six factors were each examined using varimax and oblimin rotations of the factor loading matrix. The five-factor oblimin solution (which explained 83.21% of the variance) was preferred, as it offered the best defined factor structure (see Table 1). The first factor had high loadings from target-present measures: hits on the three blocks of the PMT, hits on the MMT, and overall performance on the CFMT+. The second factor had high loadings from correct rejection scores on the three matching blocks, as well as overall scores from the CFMT+. The third and fourth factors represented hits and correct rejections, respectively, on the Crowds test; the fifth factor had a high loading from correct rejections on the MMT. A full correlation matrix is displayed in Table 2.

In sum, this analysis suggests that (a) the two target-present memory measures are related, but target-absent memory performance should be independently considered; (b) the three new blocks of the matching test are related, but target-present and target-absent trials should again be considered independently; and (c) both target-present and target-absent performance on the Crowds test is distinct from all other measures. These findings were used to create appropriate indices that assessed consistency of performance across related and unrelated measures.

3.2 | Consistency of face memory performance

Overall percentage correct on the MMT was calculated by summing hits and correct rejections. Norms for each of these measures were

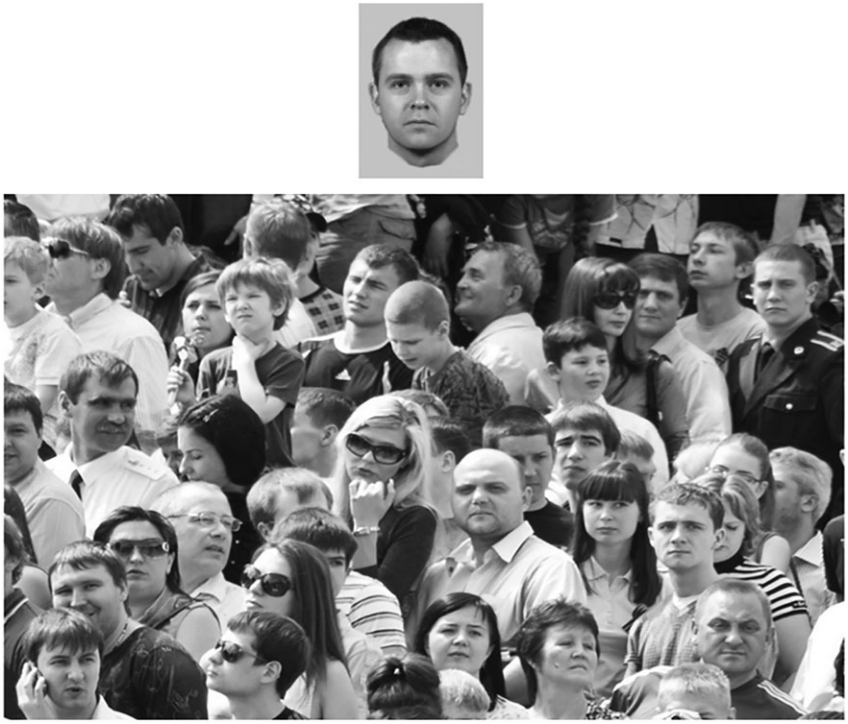

FIGURE 3 A sample target-present trial from the Crowds test

TABLE 1 Oblimin rotated component loadings for the five new face recognition tests, with separate loadings for hits and correct rejections (CRs), and the CFMT+

Component	1	2	3	4	5
CFMT+	0.50	0.55			
MMT: hits	0.70	0.37	0.30		
MMT: CRs					0.97
PMT (pose): hits	0.86				
PMT (pose): CRs		0.95			
PMT (glasses): hits	0.91				
PMT (glasses): CRs		0.75			
PMT (facial hair): hits	0.91				
PMT (facial hair): CRs		0.73			
Crowds: hits			0.98		
Crowds: CRs				0.97	

set at 1.96 *SD*s from the control mean (see Table 3). Officers' scores ranged from 53.33% to 95.56% correct, with 14 individuals exceeding the control cut-off. Eleven of these officers had also outperformed

TABLE 3 A breakdown of mean (*SD*) control performance on the MMT

	Control mean (*SD*)
Hits (%)	51.33 (20.18)
Correct rejections (%)	55.17 (23.84)
Misidentifications (%)	15.33 (11.76)
Misses (%)	33.33 (20.85)
Overall correct (%)	53.25 (14.06)
d' (sensitivity)	0.22 (0.84)
c (bias)	−0.16 (0.62)
% positive responses in TP trials that were hits (vs. misidentifications)	76.81 (15.35)

TABLE 2 Spearman's correlations for the five new face recognition tests, with separate loadings for hits and correct rejections (CRs), and the CFMT+

	CFMT+	MMT hits	MMT CRs	Pose hits	Pose CRs	Glasses hits	Glasses CRs	Facial hair hits	Facial hair CRs	Crowds hits	Crowds CRs
CFMT+	1	0.674*	0.425*	0.433*	0.532*	0.445*	0.440*	0.494*	0.517*	0.115	0.086
MMT hits		1	0.187	0.503*	0.316	0.542*	0.229	0.600*	0.334*	−0.164	0.105
MMT CRs			1	0.146	0.231	0.044	0.446*	0.061	0.254	−0.171	0.094
Pose hits				1	−0.140	0.703*	−0.002	0.713*	0.091	0.035	0.219
Pose CRs					1	−0.048	0.565*	0.089	0.607*	0.114	−0.009
Glasses hits						1	−0.155	0.739*	0.068	0.022	0.214
Glasses CRs							1	0.040	0.602*	0.080	−0.152
Facial hair hits								1	0.192	0.094	0.182
Facial hair CRs									1	0.095	−0.126
Crowds hits										1	−0.086
Crowds CRs											1

*$p < 0.005$ (Bonferroni correction applied).

controls according to the liberal inclusion criterion on the CFMT+ (nine of these also exceeded 1.96 *SD*s from the control mean), and three had not (achieving scores that were clearly within the typical range: 73.53%, 80.39%, and 80.39%). Twelve officers who had outperformed controls on the liberal CFMT+ criterion (eight surpassing the 1.96 *SD* cut-off) did not do so on the MMT, achieving scores that ranged from 64.44% to 80.00% correct (see Figure 4a).

Because the CFMT+ only contains target-present trials, we reasoned that the discrepancy in the individuals identified by overall performance on each test could result from the inclusion of target-absent trials in the MMT (as also suggested by the PCA). Thus, we examined the consistency of performance between the CFMT+ and just the hits from the MMT (see Figure 4b). Ten officers surpassed the control cut-off on MMT hits: Nine of these had outperformed controls on their overall scores for this test (the remaining individual had an overall accuracy of 78.89%), and all had also reached the liberal criterion on the CFMT+ (two failed to reach the 1.96 cut-off). Thirteen officers who had surpassed the liberal cut-off on the CFMT+ (nine of whom had surpassed the 1.96 cut-off) did not outperform controls on MMT hits. We then created an overall index of target-present face memory (Memory Hits Index: the sum of percentage hit scores on the CFMT+ and MMT) and also compared this to the single measure of target-absent face memory (correct rejections on the MMT). Eighteen officers achieved a *z* score of more than 1.96 on the Memory Hits Index: Only one of these would have been missed on the CFMT + liberal cut-off (with a *z* score of 1.17) and two different individuals according to the 1.96 cut-off. The top 10 performers on this index are displayed in Figure 4c. Notably, although index scores were mostly consistent with CFMT+ performance, there was greater variability in target-absent scores.

Given this variation in target-absent performance, there may be added value in considering correct rejections as a further performance indicator. We explored this issue using signal detection analyses and computed scores of sensitivity (d') and bias (*c*) for each individual. Information from hits and false positives were used to calculate d'—a measure of sensitivity that is free from the influence of response bias (Macmillan & Creelman, 2005). Values for the current test can range from −4.59 (consistently incorrect responding) to 4.59 (perfect accuracy), with a score of 0 indicating chance performance. Response bias is indicated by *c* and assesses whether the participant has a tendency to elicit target-present or target-absent responses (Macmillan & Creelman, 2005). Positive scores indicate more conservative responding (i.e., the tendency to make target-absent responses) whereas negative scores represent more liberal decisions (i.e., the tendency to make target-present responses); a score of 0 is a neutral response criterion. All target-present responses (i.e., hits and misidentifications) were included in this analysis, allowing us to calculate a measure of response bias that indexed a tendency to make target-present or target-absent decisions.

Because d' accounts for both target-present and target-absent performance, we examined the top performers on this measure in comparison with their identified scores on the two memory tests and the overall Memory Hits Index. Twelve officers achieved d' scores that were at least 1.96 *SD*s above the control mean (see Figure 4d). All but one (the lowest d' performer) had been identified by their overall scores on the MMT, and all but a different individual (the second poorest d' performer) on the Memory Hits Index. However, overall MMT scores identified three further individuals who did not reach criteria on d', and the Memory Hits Index identified seven additional officers. Three of the superior d' officers had not reached the 1.96

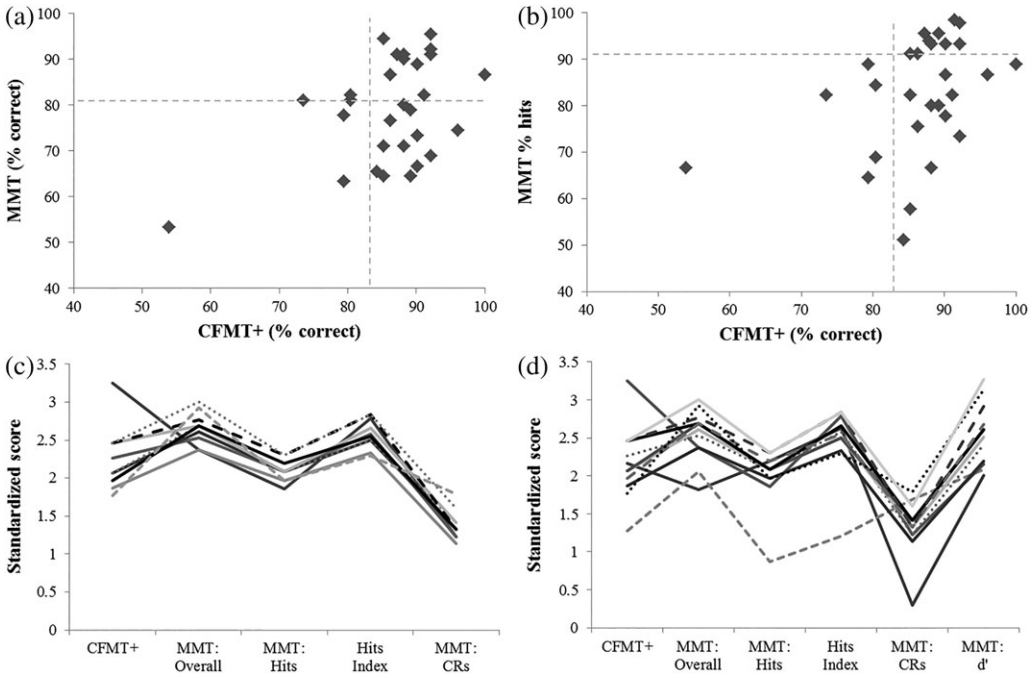

FIGURE 4 The relationship between officers' performance on the CFMT+ and (a) overall accuracy score on the MMT and (b) percentage hits on the MMT. Control cut-offs (1.5 *SD*s from the mean on the CFMT+ and 1.96 *SD*s on the MMT) are indicated by grey dashed lines. Summary of performance for (c) the top 10 performers according to the Memory Hits Index and (d) the 12 officers that surpassed control performance by at least 1.96 *SD*s on the MMT d' measure

criteria on the CFMT+ (two had surpassed the 1.5 *SD* cut-off), and eight officers who had reached the CFMT+ 1.96 cut-off were not identified by *d′*.

Next, we investigated whether the facilitated performance of the 12 superior *d′* officers resulted from differences in response bias relative to controls. No difference was observed between these officers (*M* = −0.16, *SE* = 0.09) and controls (*M* = −0.16, *SE* = 0.10) for *c*, $t(50) = 0.014$, $p = 0.989$. Further, a two-way mixed analysis of variance (ANOVA) with group (SRs and controls) and correct response type (hits and correct rejections) confirmed that, averaged across the two types of responses, SRs outperformed controls, $F(1, 50) = 74.380$, $p = 0.001$, $\eta\rho^2 = 0.598$), but there was no main effect of response type nor a significant interaction between group and the type of correct response, $F(1, 50) = 0.018$, $p = 0.894$, and $F(1, 50) = 0.793$, $p = 0.377$, respectively (see Figure 5a). In other words, the effects were not driven disproportionately by correct responses on target-present or target-absent trials. SRs also made a smaller number of misidentification errors than the control group, $t(50) = 4.187$, $p = 0.001$, $d = 1.69$; this effect held when the number of misidentifications was controlled for the number of overall positive identifications in target-present trials (by calculating the proportion of positive responses in target-present trials that were hits versus misidentifications), $t(50) = 4.908$, $p = 0.001$, $d = 1.99$ (see Figure 5b).

Thus, officers who excelled at this task showed enhanced sensitivity relative to controls, rather than a change in response bias (i.e., a general tendency to say that the target is present/absent). This conclusion is further supported by the analysis of misidentifications. Overall, SRs made less misidentification errors than controls, even when the number of misidentifications was controlled for the overall number of "target-present" responses. This indicates that the SRs were not simply guessing when they indicated that a target was present in a trial—instead, they were able to accurately identify the target faces substantially more often than control participants.

3.3 | Consistency of face matching performance

Our next set of analyses examined the consistency of performance across the three new blocks of the face matching test (i.e., the Pose, Glasses, and Facial Hair manipulations). Hits, correct rejections, and overall accuracy were summed for all participants on each block, and norms for each measure were calculated using the control data. Cut-offs were again set at 1.96 *SD*s above the control mean (see Table 4). We initially examined overall accuracy rates in each block. First, we looked at the officers who had outperformed controls in the screening version of the PMT. Of these 20 officers, 15 exceeded control performance on at least one of the three blocks: Three outperformed controls on all three blocks (see Figure 6a), nine on any two blocks (see Figure 6b), and three on any one block (see Figure 6c). Five did not outperform controls on any block (see Figure 6d). Next, we looked at the performance of the 10 officers who had not passed the initial PMT screen (i.e., they were included in this study on the basis of their CFMT+ score alone). Remarkably, only one officer failed to exceed control criterion on any one block, and two only surpassed controls on any one block (see Figure 6e). Two officers surpassed control performance on all three blocks and five on any two blocks (see Figure 6f). Overall, only five of the 30 officers showed consistently high performance across all three blocks, whereas 24 individuals surpassed criterion on any one attempt.

An alternative means of determining a cut-off for superior face matching skills is not to examine the number of tests where criterion is exceeded, but to sum all scores and index this figure against an overall criterion. However, we initially investigated the relative difficulty of the three blocks of the matching test, taking account of target-present and target-absent trials (as indicated by the PCA). Data were collapsed across all participants and entered into a 3 (block: pose, glasses, facial hair) × 2 (trial: hits, correct rejections) ANOVA. There was a significant main effect of

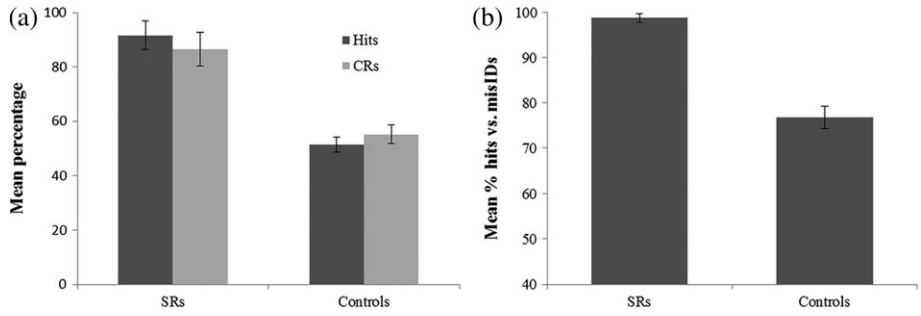

FIGURE 5 For the 12 officers who surpassed the control 1.96 *SD* cut-off on *d′*, (a) the mean percentage of hits and correct rejections on the MMT and (b) the percentage of positive responses in target-present trials that were hits (vs. misidentifications: misIDs). SR performance is displayed in relation to that of controls; error bars represent standard error

TABLE 4 A breakdown of mean (*SD*) control performance on the three new blocks of the face matching test

	Pose	Glasses	Facial hair
Hits (%)	75.10 (13.75)	64.17 (20.41)	66.98 (16.54)
Correct rejections (%)	68.44 (15.64)	81.56 (16.16)	84.90 (13.20)
Overall correct (%)	71.77 (7.66)	72.86 (8.49)	75.94 (9.06)
A (sensitivity)	0.79 (0.08)	0.82 (0.08)	0.84 (0.09)
b (bias)	0.98 (0.46)	1.73 (1.13)	1.62 (0.78)

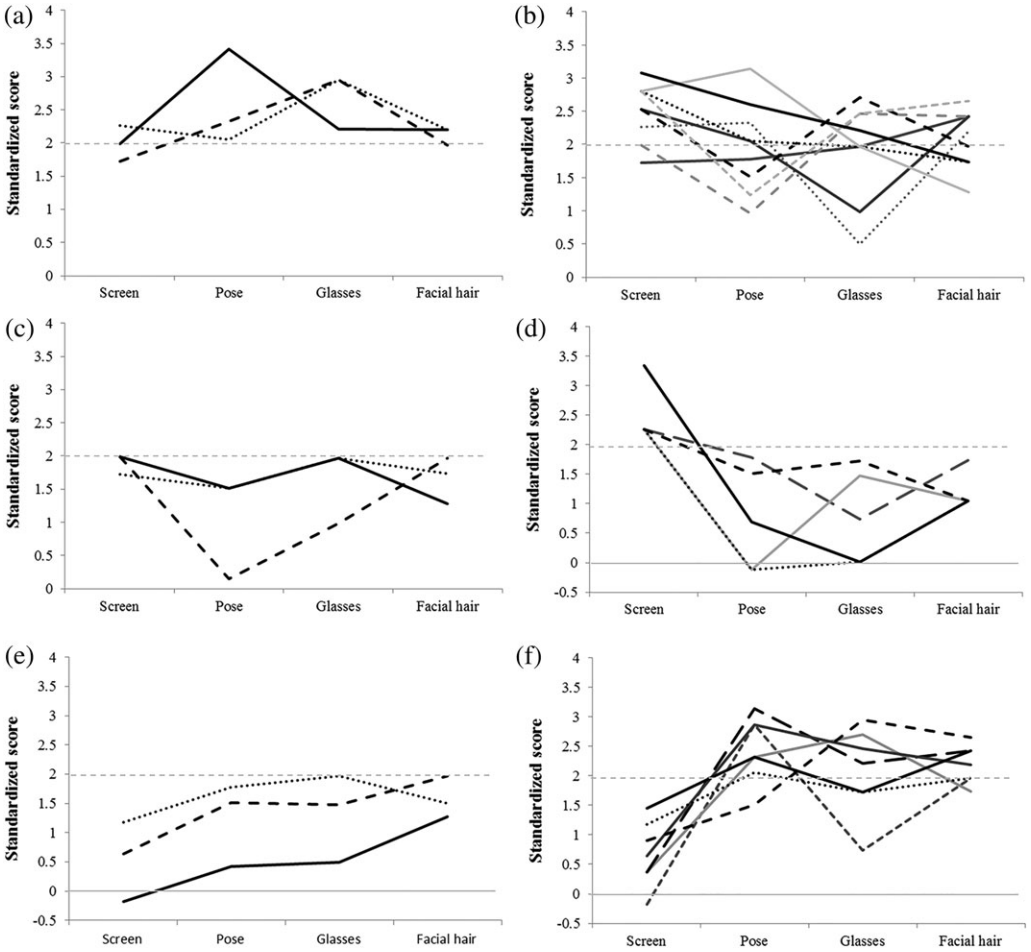

FIGURE 6 Consistency of officers' performance on the PMT at screening and in the three new blocks. Figures demonstrate those who outperformed controls at screening (according to the liberal 1.5 *SD* cut-off); then by the more conservative 1.96 *SD* cut-off on (a) all three blocks, (b) any two blocks, (c) any one block, and (d) no further block; and those who did not pass the initial screening criterion but outperformed controls on (e) only one or no block, or (f) on any two or three blocks

test, $F(2, 138) = 17.191$, $p = 0.001$, $\eta\rho^2 = .199$; follow-up analyses indicated that scores were higher in the facial hair test ($M = 83.21\%$, $SE = 1.35$) than the pose ($M = 77.59\%$, $SE = 1.20$) and glasses ($M = 79.26\%$, $SE = 1.30$) tests, with no significant difference between the latter, $F(1, 69) = 31.595$, $p = 0.001$, $\eta\rho^2 = .314$, and $F(1, 69) = 2.854$, $p = 0.096$, respectively. A main effect of trial indicated that participants made more correct rejections ($M = 83.55\%$, $SE = 1.46$) than hits ($M = 76.49\%$, $SE = 1.83$) across all tests, $F(1, 69) = 8.810$, $p = 0.004$, $\eta\rho^2 = .113$, although this was tempered by a significant interaction between the two factors, $F(2, 138) = 44.690$, $p = 0.001$, $\eta\rho^2 = 0.393$ (see Figure 7a). Specifically, participants made a larger proportion of hits in the pose test ($M = 52.34\%$, $SE = 0.99$) than the glasses ($M_{\text{proportion hits}} = 45.15\%$, $SE = 1.17$) and facial hair ($M_{\text{proportion hits}} = 45.49\%$, $SE = 0.83$) tests, $F(1, 69) = 66.943$, $p = 0.001$, $\eta\rho^2 = 0.492$. As can be seen from the mean scores, participants made a larger proportion of hits than correct rejections in the pose test, but the reverse pattern emerged in the glasses and facial hair tests. No difference in performance was observed between the latter two tests, $F(1, 69) = 0.185$, $p = 0.668$.

We then proceeded to look at overall performance across the three blocks of the test for each individual officer. Four indices of performance were created: a Matching Hits Accuracy Index (by summing the number of hits achieved on each block), a Matching Correct Rejections Accuracy Index (by summing the number of correct rejections achieved on each block), a Matching Hits Consistency Index (by calculating the variance between the number of hits achieved on each block), and a Matching Correct Rejections Consistency Index (by calculating the variance between the number of correct rejections achieved on each block).

The performance of each individual officer on each index is displayed in Figure 8a, with all four indices converted to standardized scores for ease of comparison. A correlation matrix is presented in Table 5. There were strong relationships between accuracy and consistency for both hits and correct rejections; however, although consistency of performance was related across hits and correct rejections, accuracy was not. These findings indicate that although it is important to assess accuracy of performance independently for target-present and target-absent trials, consistency is more stable. The top 10 performers on the Matching Hits Accuracy Index are displayed in Figure 8b. Only half of these individuals would have been picked up in the screening PMT, with *z* scores ranging from 0.37 to 1.72 in the remaining five officers. As observed for the memory tests, the top performers on matching hits displayed more varied performance on the target-absent trials. Notably, one

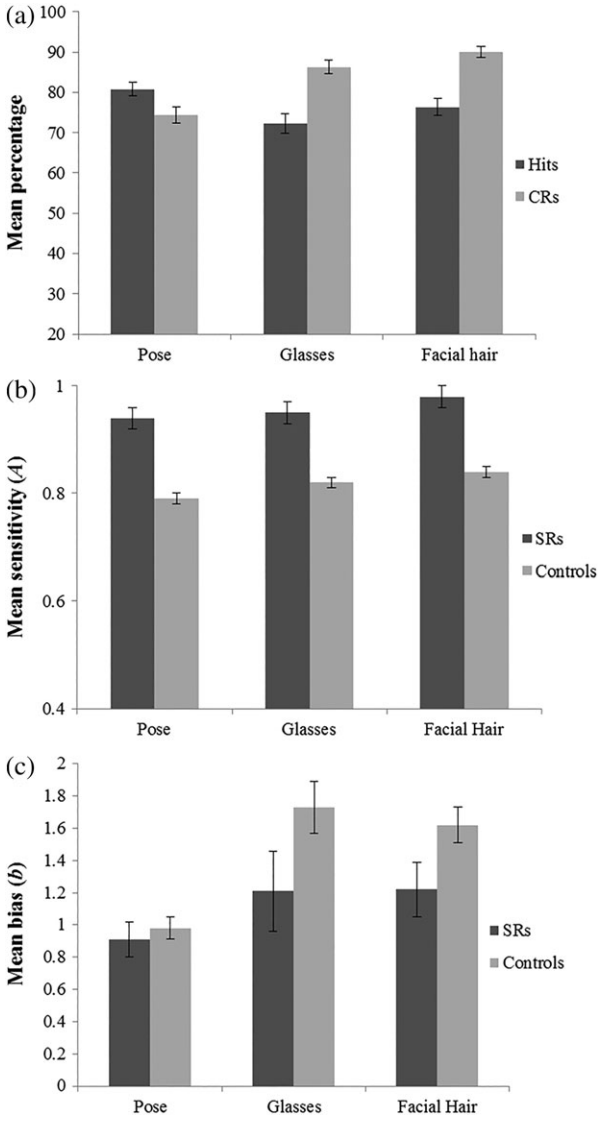

FIGURE 7 (a) Mean percentage of hits and correct rejections on each matching task across all participants and (b) mean sensitivity and (c) bias for SR officers (N = 20) and controls across the three tasks

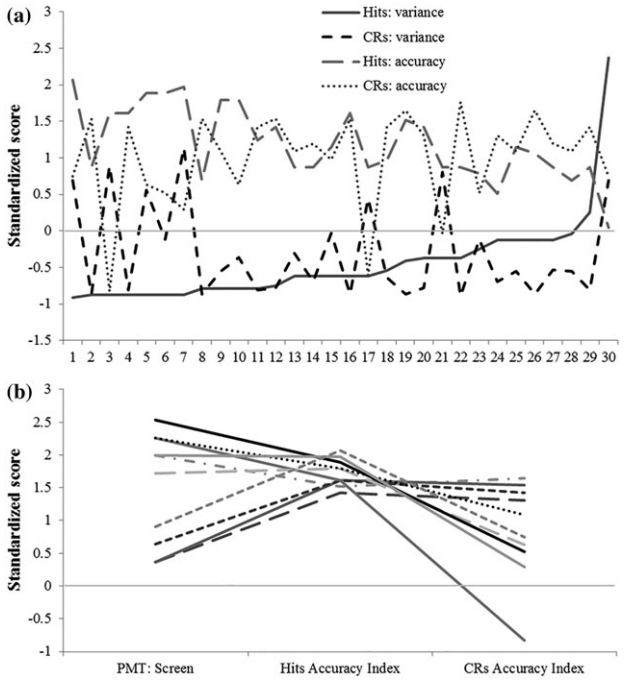

FIGURE 8 (a) The distribution of standardized scores representing the accuracy and consistency index scores of each individual officer, in terms of hits and correct rejections (CRs), across the three blocks of the matching test. Positive scores indicated high accuracy and low consistency. Officers are ordered according to the most consistent performers on hits. The top 10 performers according to the Matching Hits Accuracy Index are displayed in (b)

officer achieved a z score of −0.83 on the Matching Correct Rejections Accuracy Index.

Because of the difference in target-present and target-absent performance, we again calculated signal detection measures. As the data were not normally distributed, we used alternative, non-parametric measures of sensitivity (A) and bias (b; Zhang & Mueller, 2005). A has values that range from 0 (chance performance) to 1 (perfect performance), whereas values of b (positive vs. negative scores) have a similar interpretation to criterion c.

Mean A scores were calculated across the three blocks for each participant. Twenty officers achieved a score that was more than 1.96 SDs from the control mean (see Figure 7b). Only nine of the 20 superior performers would have been identified by a 1.96 SD cut-off on the PMT screen and a further five under the more liberal 1.5 SD cut-off. Nine officers would have been identified by their Matching Hits Accuracy Index score and 11 according to their Matching Correct Rejections Index Accuracy score. However, nine of these 11 officers were different individuals to those identified by the Matching Hits Accuracy Index officers, and this index alone would have identified one further individual who did not meet criterion on the A measure.

TABLE 5 Correlation matrix for accuracy and consistency index scores, separately for hits and correct rejections (CRs), across the three blocks of the matching test

	Hits: accuracy	Hits: variance	CRs: accuracy	CRs: variance
Hits: Accuracy	1	−0.662*	−0.167	0.207
Hits: Consistency		1	0.070	0.988*
CRs: Accuracy			1	−0.846*
CRs: Consistency				1

Note. Lower scores represent more consistent performance, whereas higher scores represent more accurate performance.

*p < 0.001 (Bonferroni correction applied).

Finally, we investigated the influence of response bias at the group level, comparing the performance of the 20 superior A officers to that of controls. A 2 (participant: SR, control) × 3 (block: pose, glasses, facial hair) ANOVA on bias (b) revealed a significant main effect of test, F (2, 116) = 17.631, p = 0.001, $\eta\rho^2$ = 0.233. Follow-up analyses confirmed more liberal responding on the pose block (M = 0.98, SE = 0.06) compared with either the glasses (M = 1.53, SE = 0.13) or facial hair (M = 1.44, SE = 0.09) blocks, F (1, 58) = 39.002, p = 0.001, $\eta\rho^2$ = 0.402, with no difference between the latter two, F (1, 58) = 0.759, p = 0.387. The main effect of group was not significant, nor did it interact with test, F (1, 58) = 2.495, p = 0.127 and F (2, 116) = 2.469, p = 0.089,

respectively (see Figure 7c). Similarly to performance on the memory tests, these results confirm that SRs excel at face matching due to better sensitivity, as opposed to a change in response bias.

3.4 | Crowds test

Hits and correct rejections were calculated for the Crowds test and summed to index overall accuracy. Controls achieved scores that ranged from 28.13% to 81.25% (see Table 6). There was no significant difference in the number of hits compared with correct rejections for controls, $t(39) = 0.189$, $p = 0.851$. Norms were once again set at 1.96 standard deviations from the control mean, yet no officer surpassed the cut-off for overall accuracy. When d' was calculated, the same pattern was observed.

These results suggest that it is difficult to surpass the 1.96 cut-off on the Crowds test—perhaps because composites constructed from memory are difficult to recognize or match to target (see discussion below). We therefore lowered the criterion and examined the performance of participants who had performed more than one *SD* above the control mean on d'. Six officers (20.00% of the sample) and eight controls (20.00% of the sample) exceeded this criterion. These individuals were combined and compared with the remainder of the control group ($N = 32$). A two-way mixed ANOVA with group (SRs and controls) and correct response type (hits and correct rejections) confirmed that, averaged across the two types of responses, the higher performers ($M = 74.58\%$ correct, $SE = 2.53$) outperformed the rest of the control sample ($M = 57.26$, SE = 1.67), $F(1, 44) = 32.642$, $p = 0.001$, $\eta\rho^2 = 0.426$; there was no main effect of response type nor a significant interaction between group and the type of correct response, $F(1, 44) = 0.053$, $p = 0.818$ and $F(1, 44) = 0.030$, $p = .864$, respectively. No difference in bias (b) was observed between the two groups, $t(44) = 0.173$, $p = 0.863$.

3.5 | Consistency of performance between unrelated measures

Finally, we used the most informative measures identified above to look at the consistency of performance across tests that tap different processes. The initial PCA permitted us to combine measures across target-present face memory, but target-absent performance also needs to be considered. We therefore used the measure that combines both types of trial: d' score on the MMT. For face matching, the PCA indicated that performance on the three new blocks of the PMT could be combined separately for target-present and target-absent trials. Although the officers demonstrated consistency in their performance across both types of trial, combined accuracy scores varied more substantially and provided a means to discriminate superior performers. We therefore selected the signal detection measure of sensitivity (A) to index overall face matching accuracy over target-present and target-absent trials. The PCA also indicated that the Crowds test was not related to the other measures, and target-present and target-absent performance should again be considered independently. Thus, we again used d' as the critical measure on this test.

We initially looked at performance across all three measures. Using a 1.96 *SD* cut-off for the memory and matching measures and a 1.00 *SD* cut-off (see above) for the Crowds test, it was found that only one officer achieved superior scores across all three indicators. We had expected that performance on the Crowds test would be related to that on the Matching test. However, although one of the two top-performing officers on the Crowds test achieved a superior score on the matching index, none of the other four top performing Crowds officers achieved a superior score on either the memory or matching indices. Thus, in line with the findings of the initial PCA, it appears that better performance on the Crowds test has little relationship with either the face memory or face matching measures.

Finally, we were theoretically motivated to look for dissociations in face memory and face matching performance. No correlation was observed between the two sensitivity measures for each type of test in the 29 officers, $r = 0.033$, $p = 0.861$ (see Figure 9). There was some evidence of an association between superior memory and matching skills on a categorical level, with nine of the 12 superior "face memorizers" also achieving "super matcher" status. However, the remaining three had z scores of 0.36, 1.08, and 1.23—suggesting that their facilitated skills are restricted to face memory. Likewise, 10 of the 19 super matchers who did not show superior memory skills showed a variety of memory z scores, with four individuals scoring below one *SD* of the control mean: 0.53, 0.71, 0.68, and 0.92. A statistical dissociation between facilitated face memory and typical face matching skills was confirmed in one officer, using Crawford and Garthwaite (2002) Bayesian Standardized Difference Test (see Table 7). Finally, it is of note that six officers did not achieve a superior score on either measure (see Figure 9). Although four of

TABLE 6 A breakdown of mean (*SD*) control performance on the Crowds test

	Control Mean (*SD*)
Hits (%)	61.72 (17.05)
Correct rejections (%)	61.09 (16.35)
Overall accuracy (%)	61.48 (12.96)
d'	0.68 (0.63)
c bias	0.06 (0.35)

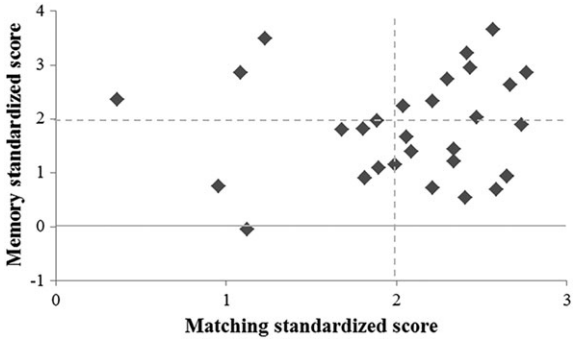

FIGURE 9 Relationship between standardized sensitivity scores on the memory (d') and matching (A) measures for the officer sample ($N = 29$)

TABLE 7 The statistical dissociation between face matching and face memory performance in one officer

Test scores		Bayesian standardized difference test: CFMT+ vs. PMT		
Memory *d′*	Matching *A*	*T*	*p**	% population more extreme
3.14	0.88	1.834	0.037	3.71

*Holm's sequential Bonferroni correction applied.

these individuals achieved scores that were close to the cut-offs, two did not (one had *z* scores of 1.12 and −0.01 for matching and memory, respectively; the other 0.96 and 0.75). The former individual had been included in the sample on the basis of their CFMT+ performance alone (which had surpassed the 1.5 but not the 1.96 *SD* cut-off) and the latter on the basis of the matching performance alone (surpassing even the 1.96 *SD* cut-off). These findings demonstrate the need for more in-depth screening protocols than are currently used in most SR investigations.

4 | DISCUSSION

This investigation aimed to investigate the consistency of superior face recognition skills both across tasks that tap the same process and between tasks that tap different processes. A sample of 30 police officers who had surpassed a liberal cut-off (1.5 *SD*s above the control mean) on a test of face memory and/or face matching took part in a battery of tests: a new face memory test that included target-present and target-absent trials, three new blocks of a face matching test (where faces differed according to pose, or the presence of glasses or facial hair), and a Crowds test that required the "spotting" of a target composite face within a crowd of faces. Results indicated that an individual's performance can vary across attempts at related tests, and superior performance does not necessarily hold across tests that tap different aspects of face processing. Critically, 30% of our sample would have been "missed" if relying solely on a CFMT+ accuracy score that is 1.96 *SD*s above the control mean, whereas another 30%, who would have been considered SRs, did not show consistently superior performance across multiple tests of face memory.

The major implication of these findings concerns current protocols for SR screening. Most published reports to date rely on criterion performance on a single attempt at the CFMT+ for inclusion in an experimental sample. However, our findings illustrate the need to examine consistency of performance across attempts at multiple related tasks. Indeed, when performance on a second measure of target-present face memory was compared with CFMT+ scores, there were some differences in the individuals who were identified as superior performers on each test. In part, this comes from the use of rather arbitrary statistical cut-offs for determining atypical performance. This issue may be overcome by creating overall index scores that can more reliably identify top performers. However, a further issue was repeatedly encountered: differences in target-present versus target-absent performance. Although there was some consistency in target-present performance across individual scores on the CFMT+ and MMT, there was much more variability in target-absent scores. This finding poses a practical problem, as different individuals tended to excel at each measure. In policing practice, the correct answer to a facial identity challenge is not known—that is, it is not possible to know whether an officer should be deployed who is particularly good at target-present trials versus one who is particularly good at target-absent performance. Perhaps the best solution is to identify the top performers on measures that encompass both types of performance, such as sensitivity scores calculated from signal detection theory. Although the top performers on these measures may not be the top performers on target-present or target-absent indices, they are the most consistent overall performers when response bias is accounted for. This is a particularly important issue in real-world face recognition scenarios such as policing, where false leads or even miscarriages of justice can result from errors in either target-absent or target-present judgments. Thus, although we agree that the CFMT+ is an excellent test of target-present face memory, it needs to be supplemented by measures of target-absent face memory to provide a full and informed assessment of top-end face memory performance.

The importance of independent assessment of target-present and target-absent performance also came through for face matching: although consistency was highly correlated across the two types of trial, accuracy was not. This finding suggests at least some stability in repeated performance at the same task, although it should be noted that the analyses were carried out on overall index scores. Although combined scores may eliminate some of the noise that present in isolated test scores, some caution may need to be exercised when creating combining performance across multiple attempts at related tasks. For instance, different patterns of response bias were noted for the "pose" matching items compared with the glasses and facial hair manipulations, perhaps because changes in viewpoint require more substantial 3D transformations than judgments on frontal faces (i.e., when glasses or facial hair are added or removed, but viewpoint does not change). Future work should explore whether different task demands return different superior performers and consequently whether overall indices should be restricted to only the most similar tasks (if the aim is to identify the best performers for specific tasks), or include a range of tasks (if the aim is to identify the most consistent overall performers).

From a theoretical perspective, it seems likely that the finding that different individuals excel at target-present versus target-absent performance results from a genuine independence between the two measures. Indeed, we found no evidence of differences in response bias between SRs and controls on any measure. Further, because SRs are operating at such a high level of sensitivity, it is extremely unlikely that response bias could explain their performance. Instead, the findings reported here fit well with previous work using typical perceivers that suggest a dissociation between target-present and target-absent performance for the matching of unfamiliar faces—an effect that gradually disappears as faces increase in familiarity (Megreya & Burton, 2007). Interestingly, the results reported here extend this finding by suggesting that the effect may hold even for top-end performers—indicating that even these individuals do not have an absolute ability to tolerate within-person variability in images of unfamiliar individuals (see Young & Burton, 2017, 2018).

The findings reported here also offer evidence for a potential dissociation between different types of superior performer at a broader level, as different patterns of facilitated face matching versus face memory skills were uncovered. This variability in SR presentation has previously been reported in small case series (e.g., Bobak et al., 2017), and a statistical dissociation between face matching and face memory for three SRs was offered in a recent publication from our laboratory (Bate et al., 2018). However, those individuals presented with superior face matching but typical face memory skills—the reverse pattern to the individual described in the current paper. This is theoretically important as previous evidence of "super-matchers" without facilitated face memory skills, but not vice versa, suggested that enhanced perceptual processes underpin facilitated face memory performance. The individual reported here suggests this is not necessarily the case. However, an important but unanswered question concerns the domain-specificity of these dissociations. Indeed, an individual with superior face memory but not matching skills might be benefiting from a more general enhancement in memory.

Finally, performance on the Crowds test deserves specific consideration. The results reported here converge with our previous work, where only one individual from a sample of 200 self-referred SRs outperformed controls on this task by more than 1.96 *SD*s (Bate et al., 2018). Given this previous study utilised a civilian SR sample, we questioned whether the inherent difficulty of dealing with error within facial composites (particularly those that have been constructed from memory) may have constrained performance and therefore whether a sample of SR police officers (who likely have more experience with composite stimuli) may perform better on this task. The current findings suggest this is not the case, although we did not explicitly enquire about experience with facial composites when collecting our data. Nevertheless, exactly the same proportion of officers and control participants surpassed the 1.0 *SD* criterion on the Crowds test—a pattern that did not emerge on the other tests, where only one control participant surpassed the 1.96 cut-off on the Memory d' index, and no controls exceeded the same cut-off on the Matching *A* index.

Thus, it may be that composite face recognition tasks are difficult even for SRs who have at least some familiarity with this type of artificial image. Indeed, inaccuracies in the shape and appearance of individual features on composite stimuli, in addition to their spatial positioning (e.g., Frowd et al., 2005), can result even from protocols that are designed to create identifiable images (e.g., Frowd et al., 2012). Consequently, such composite faces are usually much harder to recognize, or even to match to target, than photographs of the target identities themselves (e.g., Frowd et al., 2014; Frowd, Bruce, McIntyre, & Hancock, 2007). These inaccuracies in the size, shape, and positioning of features may be what disrupts the performance of SRs on the Crowds task: SRs may be exceptional at recognizing the highly stable properties of faces (as tapped in tests such as the CFMT+, which has highly controlled images), but relatively less adept at spotting more general "likenesses" between faces.

This hypothesis is supported by the overall patterns of performance observed here. Although the composite faces used in the Crowds test present the most challenging instances of facial variability in the current battery of tasks, it is pertinent that both the MMT and matching tasks used more ambient facial images than the CFMT+. In both this study and that reported by Bate et al. (2018), the MMT appears more sensitive to top-end performance than the CFMT+ (see also Bate et al., 2018)—discriminating between individual SRs who achieved very similar scores on the latter test. Likewise, the finding that some SRs can excel at the matching tests but not the Crowds test (and not vice versa) may also be explained by the relative difference in within-person variability between these two tasks. Thus, it may be that the ability to complete more challenging face recognition tasks reflects properties of the images themselves, rather than different individuals being suited to different tasks. In any case, wider screening of personnel using tasks that directly replicate real-world needs should be initiated (see Balsdon, Summersby, Kemp, & White, 2018), and future work might examine the limits of super recognition with regard to image variability.

In sum, the above discussion indicates that (a) task demands of screening tests need to be thoroughly assessed prior to implementation, (b) multiple assessments should be carried out and index scores calculated, (c) screening should allow for different individuals to be short-listed for different tasks, and (d) the best overall performers will likely not be those that excel on target-present measures alone. Although signal detection measures may offer the best indices of all-round performance, the use of any particular statistical cut-off alongside these measures only offers an arbitrary means of identifying SRs. What may work best in practice is to rank personnel on their overall performance, calculated from multiple attempts at specific tests containing target-present and target-absent items, to create a "leader board" for each required task. At any point in time, the best available personnel may then be selected for a particular task in hand.

ACKNOWLEDGEMENT

S. B. is supported by a British Academy Mid-Career Fellowship (MD170004).

ORCID

Sarah Bate https://orcid.org/0000-0001-5484-8195

REFERENCES

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, *3*, 25. https://doi.org/10.1186/s41235-018-0114-7

Bate, S., & Bennetts, R. (2015). The independence of expression and identity in face-processing: Evidence from neuropsychological case studies. *Frontiers in Psychology*, *6*, 770. https://doi.org/10.3389/fpsyg.2015.00770

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

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

Bate, S., Haslam, C., Tree, J. J., & Hodgson, T. L. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex*, *44*, 806–819. https://doi.org/10.1016/j.cortex.2007.02.004

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

Bennetts, R. J., Mole, J. A., & Bate, S. (2017). Super recognition in development: A case study of an adolescent with extraordinary face recognition skills. *Cognitive Neuropsychology*, *34*, 357–376. https://doi.org/10.1080/02643294.2017.1402755

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*, 277–291. https://doi.org/10.1037/a0029635

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

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. https://doi.org/10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS One*, *11*, e0148148. https://doi.org/10.1371/journal.pone.0148148

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

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

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*, 339–360.

Burton, A. M., White, D., & McNeil, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*, 286–291. https://doi.org/10.3758/BRM.42.1.286

Chatterjee, G., & Nakayama, K. (2012). Normal facial age and gender perception in developmental prosopagnosia. *Cognitive Neuropsychology*, *29*, 482–502. https://doi.org/10.1080/02643294.2012.756809

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, *40*, 1196–1208. https://doi.org/10.1016/S0028-3932(01)00224-X

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*, 419–430. https://doi.org/10.1080/02643290701380491

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 subjects. *Neuropsychologia*, *44*, 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Fodarella, C., Kuivaniemi-Smith, H., Gawrylowicz, J., & Frowd, C. D. (2015). Forensic procedures for facial-composite construction. *Journal of Forensic Practice*, *17*, 259–270. https://doi.org/10.1108/JFP-10-2014-0033

Frowd, C. D., Bruce, V., McIntyre, A., & Hancock, P. J. B. (2007). The relative importance of external and internal features of facial composites. *British Journal of Psychology*, *98*, 61–77. https://doi.org/10.1348/000712606X104481

Frowd, C. D., Carson, D., Ness, H., Richardson, J., Morrison, L., McLanaghan, S., & Hancock, P. J. B. (2005). A forensically valid comparison of facial composite systems. *Psychology, Crime & Law*, *11*, 33–52. https://doi.org/10.1080/10683160310001634313

Frowd, C. D., Skelton, F. C., Atherton, C., Pitchford, M., Hepton, G., Holden, L., … Hancock, P. J. B. (2012). Recovering faces from memory: The distracting influence of external facial features. *Journal of Experimental Psychology: Applied*, *18*, 224–238. https://doi.org/10.1037/a0027393

Frowd, C. D., White, D., Kemp, R. I., Jenkins, R., Nawaz, K., & Herold, K. (2014). Constructing faces from memory: the impact of image likeness and prototypical representations. *Journal of Forensic Practice*, *16*, 243–256. https://doi.org/10.1108/JFP-08-2013-0042

Jenkins, R., White, D., Van Montford, X., & Burton, A. M. (2011). Variability in photos of the same person. *Cognition*, *121*, 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*, 841–845. https://doi.org/10.1002/acp.3261

Lee, Y., Duchaine, B., Wilson, H. R., & Nakayama, K. (2010). Three cases of developmental prosopagnosia from one family: Detailed neuropsychological and psychophysical investigation of face processing. *Cortex*, *46*, 949–964. https://doi.org/10.1016/j.cortex.2009.07.012

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

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

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R. B., Rivolta, D., … O'Connor, K. B. (2011). Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test-Australian. *Cognitive Neuropsychology*, *28*, 109–146. https://doi.org/10.1080/02643294.2011.616880

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*, 1175–1184. https://doi.org/10.3758/BF03193954

Murray, E., Hills, P. J., Bennetts, R. J., & Bate, S. (2018). Identifying hallmarks symptoms of developmental prosopagnosia for non-experts. *Scientific Reports*, *8*, 1690. https://doi.org/10.1038/s41598-018-20089-7

Ramon, M., Miellet, S., Dzieciol, A. M., Konrad, B. N., Dresler, M., & Caldara, R. (2016). Super-memorizers are not super-recognizers. *PLoS One*, *11*, e0150972. https://doi.org/10.1371/journal.pone.0150972

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police Super-Recognisers. *PLoS One*, *11*, e0150036. https://doi.org/10.1371/journal.pone.0150036

Russell, R., Chatterjee, G., & Nakayama, K. (2012). Developmental prosopagnosia and super-recognition: No special role for surface reflectance processing. *Neuropsychologia*, *50*, 334–340. https://doi.org/10.1016/j.neuropsychologia.2011.12.004

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

Young, A. W., & Burton, A. M. (2017). Recognizing faces. *Current Directions in Psychological Science*, *26*, 212–217. https://doi.org/10.1177/0963721416688114

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

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212. https://doi.org/10.1007/s11336-003-1119-8

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Bate S, Frowd C, Bennetts R, et al. The consistency of superior face recognition skills in police officers. *Appl Cognit Psychol*. 2019;1–15. https://doi.org/10.1002/acp.3525