Searching for faces in crowd chokepoint videos

Robin S. S. Kramer, Sarah C. Hardy, and Kay L. Ritchie

School of Psychology, University of Lincoln, UK

Corresponding Author:
Robin Kramer, School of Psychology, University of Lincoln, Lincoln LN6 7TS, UK.
E-mail: remarknibor@gmail.com

Telephone: +44 (0)1522 835806

Running Head: Searching for faces in crowd videos

Conflict of Interest

The authors have no conflict of interest to declare.

Acknowledgements

The authors thank Joanne Prior, Amberley Westerman, Tom Bayer, and Magdalena Zajaczkowska for their help with data collection. We also thank Peter Hancock for the idea of using a lecture theatre exit for filming crowd stimuli, and Mike Burton and Andy Young for their input early on in the project. We thank Alex Jones for his help with the use of Gorilla for data collection. Finally, we thank Ferenc Igali and Foivos Vantzos for technical assistance regarding filming.

This work was supported by an Experimental Psychology Society's Small Grant awarded to R.S.S.K.

Abstract

Investigations of face identification have typically focussed on matching faces to photographic IDs. Few researchers have considered the task of searching for a face in a crowd. In Experiment 1, we created the *Chokepoint Search Test* to simulate real-time search for a target. Performance on this test was poor (39% accuracy) and showed moderate associations with tests of face matching and memory. In addition, trial-level confidence predicted accuracy, and for those participants who were previously familiar with one or more targets, higher familiarity was associated with increased accuracy. In Experiment 2, we found improvements in performance on the test when three recent images of the target, but not three social media images, were displayed during searches. Taken together, our results highlight the difficulties inherent in real-time searching for faces, with important implications for those security personnel who carry out this task on a daily basis.

Keywords

CCTV, face matching, face recognition, crowd search, chokepoint

1 Introduction

When required to confirm the identity of an individual, officials are typically faced with a comparison. For example, a submitted photograph attached to a passport application is compared with previous images of the person stored on file, or a photographic ID is compared with the person standing before them. Importantly, in the majority of cases, these types of matching tasks involve people who are unfamiliar to the viewer. A number of studies have now demonstrated the difficulties associated with both 'photo to photo' (e.g., Burton, White, & McNeill, 2010; Kramer, Mohamed, & Hardy, 2019; Kramer, Mulgrew, & Reynolds, 2018; Megreya & Burton, 2006; Megreya, Sandford, & Burton, 2013; Ritchie et al., 2015) and 'live person to photo' matching (e.g., Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008; Kramer, Mireku, Flack, & Ritchie, 2019; Ritchie, Mireku, & Kramer, 2019; White, Kemp, Jenkins, Matheson, & Burton, 2014).

Officials are also routinely required to identify unfamiliar people in crowd situations. Unlike the matching contexts described above, searching for people in crowds represents a less controlled, more complex scenario. For each passer-by, an officer must decide whether the individual's face matches the photograph of the person of interest with which she or he has been provided. However, unlike with the presentation of an ID document at border control, for instance, the officer is unable to stop every person and carry out a comparison in relatively controlled conditions. Examples of this type of task include being on the lookout for people who have previously committed offenses at sporting events or searching for known criminals in a live CCTV feed. Interestingly, in some situations, particular security officers may actually be deployed specifically because they are familiar with known troublemakers at a given sports grounds, for example. Such instances highlight the presumed benefits of familiarity during searching in crowds.

1.1 Searching in crowds without time constraints

Although searching for people in crowds is an everyday occurrence and a common feature of security protocols, few researchers have investigated this task and its likely difficulties. Bate and colleagues (2018) developed the Crowds Matching Test, where participants were presented with static photographs of crowds (downloaded from the Internet), along with target face composites constructed using the EvoFIT system.[1] Half of the 32 trials contained the target identity (target-present) and half did not (target-absent), with participants simply responding 'present' or 'absent'. Typical performance on this task was poor, with an average proportion correct of 0.63. However, it is important to note that the requirements for this task included searching and matching, but without the inherent time constraints and other complications imposed by a dynamic crowd continuously passing by the viewer.

To this end, Davis and colleagues created the Spot the Face in a Crowd Test (SFCT – Davis, Forrest, Treml, & Jansari, 2018; Durova, Dimou, Litos, Daras, & Davis, 2017) in order to explore performance in a realistic search task. Using video footage recorded at London tourist locations, eleven short clips (approximately 1-2 mins each) were produced. Participants were required to view these clips as if searching for missing people, and were provided with four photographs of each person (taken from social media) to aid their search. While reviewing these videos, each participant was given either two, four, or eight actors to search for. Actors appeared in either one or two clips, and each clip contained two, one, or zero actors. Importantly, participants were able to rewind and pause the video clips as needed, removing time constraints and mirroring real-world searches through pre-recorded CCTV footage.

As the number of actors that participants were required to search for increased, overall performance on the SFCT decreased (Davis et al., 2018). For instance, the mean proportion of hits fell from 0.77 (two actors) to 0.59 (eight actors), while the mean proportion of correct rejections (0.66 and 0.68 respectively) remained unchanged. Interestingly, police experts outperformed

[1] These composites were the result of 'evolving' face images of the targets from memory and were not themselves photographs of the target identities.

untrained, inexperienced participants on this task, as well as on a test of face memory, suggesting that these abilities may be related. It is also worth noting that familiarisation with the actors beforehand (up to 20 mins of discussing and rating the actors' perceived personalities based on their photos) resulted in performance improvements.

Finally, Mileva and Burton (2019) investigated searching within crowds using CCTV footage of a busy rail station. As with the SFCT, participants could rewind and pause the videos during the task. Again, performance was highly error-prone, with percentage correct accuracies at 67% on target-present trials and 52% when the target was absent. In addition, the researchers were able to improve performance by providing participants with multiple photographs of the target, rather than only one, during search.

While both the SFCT and Mileva and Burton's (2019) task are aimed at replicating the reviewing process of police CCTV footage, it is likely that searching for faces in crowds in real time (i.e., without the ability to pause, rewind, and replay) represents a somewhat different (though overlapping) set of task requirements.

1.2 Individual differences across tasks

For several years, researchers have been interested in the requirements involved in face memory, matching, and processing more generally. Is there some generic, underlying ability with faces that results in good performance across all tests (e.g., the factor *f* – Verhallen et al., 2017) or do different aspects of face processing require different abilities? Recent research suggests that there is some degree of commonality in the underlying mechanisms recruited across different face-identity tasks (McCaffery, Robertson, Young, & Burton, 2018). For example, measures of face matching and memory show moderate to large correlations (r = .45 to .50 – McCaffery et al., 2018; r = .48 – Verhallen et al., 2017). However, these relationships are far from perfect, suggesting the presence of independent requirements as well.

Little is known about the underlying mechanisms involved in searching for faces. Davis and colleagues (2018) reported significant correlations between face memory abilities (using the CFMT+, an extended version of the CFMT; Russell, Duchaine, & Nakayama, 2009) and their SFCT's hits (r = .18) and correct rejections (r = .17). However, the small association between face memory and search performance may be the result of allowing participants to compare photos to the paused video footage during the search process, perhaps decreasing some of the memory demands. As such, we suggest that searching for faces in real time (i.e., without the option of pausing the video) might utilise a larger memory component since viewers will likely try to minimise the amount of time spent attending to the photograph in order to avoid missing the target passing by in the crowd.

Evidence also suggests that personality may play a role in face-related tasks, although results appear to be mixed. Face recognition performance was found to be higher in extraverted individuals (Lander & Poyarekar, 2015; Li et al., 2010) although face matching does not seem to be associated with personality traits other than perhaps facets of neuroticism (anxiety – Megreya & Bindemann, 2013; no associations – Lander & Poyarekar, 2015). Indeed, a recent study found no relationship between personality factors and measures of face memory and matching (McCaffery et al., 2018). In addition, Davis and colleagues (2018) found no relationship between personality factors and performance on their SFCT task. Therefore, it remains unclear as to whether personality differences across individuals might be associated with *real-time* searching for faces.

1.3 Confidence

The relationship between confidence and performance is not well understood. Evidence has shown that people's self-rated abilities with face matching and memory may be predictive of their actual abilities, although this association may only be modest (Bate & Dudfield, 2019; Bobak, Mileva, & Hancock, 2019; Bobak, Pampoulov, & Bate, 2016; Gray, Bird, & Cook, 2017; Palermo et al., 2017;

Shah, Sowden, Gaule, Catmur, & Bird, 2015). However, research has shown that confidence at the level of individual responses does appear to reflect accuracy. For example, judges were more confident on trials in which they responded correctly in both ten-image array and same/different face matching tasks (Bruce et al., 1999; Hopkins & Lyle, 2019).

In terms of searching for faces in crowds, judges' proportions of hits were associated with their confidence on target-present trials while no such relationship was found for false positives and confidence (Davis et al., 2018). Therefore, judges may have some insight into their face searching abilities although further work is needed to confirm this.

1.4 Improving performance through variability

A key difficulty with both matching and learning/recognising unfamiliar faces is that viewers have limited or no knowledge regarding how these faces can vary across images or in real life (Jenkins, White, Van Montfort, & Burton, 2011). Given that this variability is idiosyncratic (Burton, Kramer, Ritchie, & Jenkins, 2016), knowledge of how familiar faces vary provides minimal information regarding variation for a new face (Kramer, Young, & Burton, 2018).

Recent evidence suggests that providing multiple images of an unfamiliar face may result in improvements with both matching (White, Burton, Jenkins, & Kemp, 2014; cf. Ritchie et al., 2019) and learning (Ritchie & Burton, 2017). Further, larger variation across images of the same individual appears to produce increased performance in both cases (Menon, White, & Kemp, 2015; Ritchie & Burton, 2017). As such, viewing multiple images of a target may also prove beneficial when searching for that person in a crowd. Indeed, Mileva and Burton (2019) provide initial support for this idea, finding performance gains of around 10% when either three or 16 images of the target (with no apparent difference between these conditions) were displayed in comparison with only one.

1.4 The current research

In the experiments presented here, we investigate the task of searching for unfamiliar and familiar faces in real time. We develop a test that closely mirrors this task within the laboratory, and consider its association with tests of face memory and matching. We also explore how personality, familiarity, and confidence might relate to performance in this test. Finally, we investigate whether variability information (through multiple images) regarding the target face improves search performance.

Here, we focus on a specific scenario in which crowds pass through a doorway, forming a natural "chokepoint". These types of situations are often utilised by security personnel because the temporarily narrowed crowd allows for easier inspection of all individuals as they pass by (either individually or in pairs), in comparisons with an unrestricted crowd that may appear as tens of people across its front. By posting officers at the entrances to sporting events, for example, one can increase the likelihood that most, if not all, faces will be inspected.

2 Experiment 1 – Search performance

There is currently no research investigating the task of searching for people in real time within dynamic crowds. Here, we develop the Chokepoint Search Test (CST), designed to simulate searching at live events for specific targets, and consider how performance on this test may be related to other measures of face processing, as well as personality.

2.1 Method

2.1.1 Participants

For clarity in this experiment, we labelled our participants as either targets or judges. Targets were those individuals who appeared in the video and photograph stimuli presented. Judges, in contrast, were those who participated in the CST, providing the current data for analysis.

Targets: Twenty White, female undergraduate students acted as models (age $M = 20.05$ years, $SD = 0.83$ years). This represented a subset of 50 students who volunteered and received £5 Amazon gift cards as compensation.

Judges: Ninety-seven students (85 women; age $M = 20.25$ years, $SD = 4.11$; 94% self-reported ethnicity as White) at the university took part in exchange for course credits or monetary compensation. The data from one additional judge was lost due to a technical error. There was no overlap between this sample and the 50 targets who took part in stimulus creation.

All targets and judges provided written informed consent before taking part, and received both written and verbal debriefings at the end of the experiment. The University of Lincoln's School of Psychology ethics committee approved both the creation of photograph and video stimuli, as well as the collection of behavioural data. These were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

2.1.2 Stimuli

2.1.2.1 Chokepoint videos

Videos were filmed at a lecture theatre on campus at the university. Of the three exits to the room, only one was recorded as lectures finished (with clear signage identifying this exit). We explained to the students on the course (a second-year undergraduate psychology module) that, should they choose to leave the room through the filmed exit, they were consenting to appear in the videos and that these would be used as stimuli in a face search task. There were 284 students (42 men) taking the course, although fewer left through the filmed exit in each session. The same course (and hence

the same group of students) was filmed over the semester, resulting in the collection of a separate chokepoint video for each of 17 lectures (spanning 12 weeks), with many of the students appearing in multiple videos. One of these videos was subsequently discarded due to an error during filming. For the remaining 16 videos, the number of people appearing in each ranged from 56 to 88 (M = 69.1).

The designated exit was filmed using a GoPro HERO5 Session camera mounted on a tripod (at eye level – approximately 160 cm high) outside the room, on a landing at the top of the staircase used to leave the lecture theatre. Two additional cameras were mounted inside the theatre above the exit, although these were not utilised for test construction. All three cameras recorded in colour at a resolution of 1920 x 1080 pixels, at 30 frames per second.

The videos were processed using Adobe Premiere Pro CC 2018 software. All videos were trimmed so that each one featured only a few seconds of inactivity (where no one appeared onscreen) before and after the crowd exited the lecture theatre, with the final durations ranging from 80 s to 189 s (M = 129.9 s). In addition, the frame size was cropped to remove extraneous surroundings, resulting in a resolution of 760 x 1080 pixels. The sound was also removed from the videos.

2.1.2.2 Target photographs

All students registered on the course were invited to appear as targets (to be searched for), with the proviso that they had appeared in at least one chokepoint video. As such, targets were recruited from the course after attending at least one filmed lecture and leaving through the appropriate exit. Targets' appearances within the crowds while leaving the lecture theatre were entirely unconstrained and were simply the result of their individual behaviours as the lecture finished.

Each target attended an additional session in which various materials were collected: multiple photographs were taken, social media images were downloaded, and a short conversation was filmed. Demographic information (age, gender, ethnicity) was also obtained.

For test construction, we took a colour, passport-style photograph (e.g., posing with a neutral expression) using a GoPro HERO5 Session camera, which was placed at a distance of approximately 50 cm from the model. Height was adjusted for each individual. Image resolution was 3648 x 2736 pixels. Using a custom MATLAB script, the images were resized so that all faces featured the same interpupillary distance and were cropped to include only the head and top of the shoulders. The final resolution was 380 x 570 pixels.

From the initial set of 50 targets, we selected 20 White women for use in the task. This homogeneous group was chosen because the course cohort comprised predominantly White women, and so searching for male or non-White models would have been a significantly easier task. In addition, we made sure to exclude our experimenters, who also appeared in the original set of 50 models.

During this photographic session, targets were also required to identify themselves in one or more of the chokepoint videos. Following this, two of the authors (both of whom were familiar with all targets, with one being a student in the same cohort and attending the same lecture series) identified the targets in the remaining videos. Together, this provided a 'ground truth' for the test and its subsequent analysis.

2.1.3 Procedure

Judges first completed the CST, followed by three additional measures. All judges completed all four tasks with the exception of one judge, who failed to complete the CFMT due to a technical issue.

2.1.3.1 Chokepoint Search Test

To create the CST, a single target was presented on each of 20 trials, with half randomly assigned to appear in 'target present' trials (the target is present in the video) and half in 'target absent' trials (the target is not present). We chose to present equal proportions of the two trial types for ease of analysis and interpretation, although we acknowledge that in real-world searches, the likelihood of a target (e.g., a known terrorist) being present would be very small. Previous research with face matching suggests that the low prevalence of targets may affect performance levels (Papesh & Goldinger, 2014; cf. Bindemann, Avetisyan, & Blackwell, 2010), while this has yet to be tested with searching tasks.

Given that our database included sixteen videos, we were able to feature every video once in our test, with four videos appearing for a second time (determined by which featured the remaining selected targets). Importantly, although these four videos appeared twice, every trial featured a different target. For 'target present' trials, videos were selected so that the target in question never appeared earlier than 17 s into the video (although judges were not told this) in order to give judges sufficient time to ready themselves for responding. Videos varied in length from 1 min 20 s to 3 min 9 s due to the variation in how long the students took to leave the room after each lecture.

We randomised the order of the 20 trials with the proviso that a) no two consecutive trials featured the same video; and b) both the first and second half of the test contained an equal number of target present (five) and absent trials (five). This trial order was then fixed and used for all judges. Prior to completing the test itself, judges were given a practice trial (featuring an image of 'Captain America' as the target and a 14 s video clip from the 'Captain America: Civil War' movie) in order to familiarise them with the procedure. We also reminded judges both verbally and onscreen that on some trials, the target would be absent.

On each trial, a photograph of the target appeared on the left of the screen. Judges could study this image for an unlimited amount of time before clicking with the mouse to start playing the

video. During playback, judges were instructed to click the 'freeze' button onscreen if they spotted the target. This response paused the video and required that judges draw a box around the target's head within the video frame (with this frame/box stored by the computer for later use in determining the accuracy of their response; see Figure 1). Finally, judges rated onscreen how confident they were in their selection, using a 1 (very low) to 5 (very high) Likert scale. Both the photograph of the target and the frozen video remained onscreen during this process. Upon providing this rating, the next trial would begin by displaying the next target's image onscreen.

If judges did not click 'freeze' during playback, the video would continue playing until it finished. At this point, judges would also complete a confidence rating, this time representing how confident they were that the target was absent. After providing this rating, the next trial would then begin.

It is important to note that judges were unable to pause, rewind, or skip ahead during video playback. This procedure 1) provided an upper limit on the test's duration for practical reasons, in contrast with an unlimited duration if video exploration were possible; and 2) better represented the real-world scenario in which someone is searching in a 'live' crowd and only gets one chance to spot the target (versus CCTV footage where exploration would be possible, as in the SFCT described earlier).

After completing the test, judges were asked if they recognised any of the targets prior to participation in the test. For those who did, a rating of prior familiarity was collected for each of the 20 targets through an additional onscreen task. Each image was presented, one at a time, and judges provided familiarity ratings utilising a 0 (unfamiliar) to 9 (very familiar) Likert scale, using the mouse to make their responses.

2.1.3.2 Additional measures

Upon completion of the CST, judges then completed the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006). Judges were familiarised with six target faces and were subsequently asked to identify these faces from target-present arrays consisting of three faces. Over 72 trials, this test provided a measure of an individual's face memory ability.

Next judges completed the Glasgow Face Matching Test (GFMT short version; Burton et al., 2010). On each of 40 trials, judges were shown a pair of face photographs and were asked to decide whether the two images showed the same person (20 trials) or two different people (20 trials). For 'same' trials, the two images were taken with different cameras and therefore avoided simple image matching. This test provided a measure of an individual's face matching ability, with negligible memory requirements.

Finally, judges completed the Mini-IPIP (International Personality Item Pool; Donnellan, Oswald, Baird, & Lucas, 2006). A 20-item short form of the 50-item International Personality Item Pool (Goldberg, 1999) measured five domains: neuroticism, extraversion, conscientiousness, agreeableness, and intellect/imagination. Judges rated each item on a five-point scale, ranging from 1 (very inaccurate) to 5 (very accurate), in order to quantify how well each statement described them. This questionnaire provided a measure of an individual's Big Five personality traits.

2.2 Results

2.2.1 Unfamiliar performance on the CST

Response accuracy was determined through comparison of the selected target within the saved video frame (see Figure 1) with the correct target (previously established by the target themselves). In some instances, judges had drawn empty boxes onscreen (demonstrating their intent to select someone but not those currently visible) because they had frozen the video just after the intended target had disappeared from view (due to the short time in which targets were visible). For these

responses, we referred back to the videos to see whether the correct target had indeed been present in the seconds before the screenshot was taken. If the target had been present *and* no distractor individuals had passed by in between the target and the selection, then this response was considered correct.

First, we consider only those judges who responded that they had no prior familiarity with any of the targets during the familiarity check after the CST ($N = 62$), providing a measure of unfamiliar performance. For target-present (TP) trials, responses were coded as hits (the model was identified), misidentifications (an incorrect person was identified), and misses (no person was identified). For target-absent (TA) trials, responses were coded as correct rejections (correctly making no identification) and false positives (a person was identified). These measures are summarised in Table 1, demonstrating poor performance on the test ($M = 38.7\%$). For an illustration of the large individual differences found on this test, see the Supporting Information.

In order to investigate whether motivation may have played a role in CST performance, we considered how long judges took to respond when giving false positives. If certain individuals were unmotivated, we would predict that they might respond quickly in order to end the trial and therefore complete the test sooner. In general, response times ($M = 55.0$s, $SD = 30.6$s) did not suggest that people were responding quickly so as to skip through the trials. For each judge, we calculated their average response time for false positives, and subsequently correlated these times with their overall performance on the CST. Evidence of a relationship would support the prediction that judges who performed poorly overall (as would be expected for unmotivated individuals) were also those that responded quickly, perhaps in order to finish sooner. However, we found no association between these two measures, $r(60) = -.01$, $p = .930$. As further evidence that motivation did not play a role, these response times also showed no associations with performance on the GFMT, $r(60) = .10$, $p = .463$, or CFMT, $r(59) = .03$, $p = .809$.

2.2.2 Unfamiliar performance and confidence

For each trial, judges provided their confidence regarding their response. In order to determine whether confidence ratings predicted trial accuracy, the data were analysed using a generalised linear mixed-effects model with crossed random effects (judges and trials) because each judge responded to the same twenty trials. Therefore, judges and trials variance were considered at Level 2 and residual variance at Level 1. In terms of the dataset, each judge by trial observation was the unit of analysis, with each row of data indicating the confidence rating given by that judge to that trial, the type of trial (present/absent), and the accuracy (correct/incorrect).

The fixed effects were the intercept, the effect of the confidence rating, the effect of the trial type, and the confidence x trial type interaction. Only the intercept in this model varied randomly across trials, whereas the intercept and the slopes of the confidence rating, the trial type, and their interaction varied randomly across judges. The confidence ratings given by judges were group mean centred in order to avoid conflating lower level (within-judge) and higher level (between-judge) variance.

Modelling was carried out in R 3.5.2 (R Development Core Team, 2016) using the glmer function from the *lme4* package 1.1-20 (Bates, Maechler, Bolker, & Walker, 2015). Focussing on the fixed effects, we found that confidence significantly predicted accuracy, $b = 0.52$, $SE = 0.12$, $p < .001$, with higher confidence ratings associated with correct responses. In addition, trial type also predicted accuracy, $b = -0.88$, $SE = 0.43$, $p = .039$, with higher accuracy in target-absent trials (see Table 1). Finally, the confidence x trial type interaction was significant, $b = 0.59$, $SE = 0.18$, $p < .001$, with a stronger association between confidence and accuracy in target-present trials (although a significant association was present for both trial types).

2.2.3 Unfamiliar performance and other measures

Performance on our test demonstrated large individual differences (overall percentage correct ranged from 0% to 70%) that may reflect more general face matching or memory abilities, as well as personality differences across our sample. We therefore examined the association between levels of performance on the CST and our three additional measures.

Performance on the two additional face tests is summarised in Table 1, and the relationships between performance on the CST and the additional face and personality measures are summarised in Table 2. The levels of performance on both the GFMT and CFMT found here are in line with published norms and previous research (Burton et al., 2010; Duchaine & Nakayama, 2006; McCaffery et al., 2018).

We found a significant association between CST hits and performance on the CFMT, $r(59) = .26$, $p = .046$, and the GFMT, $r(60) = .29$, $p = .025$. We also found a moderate and significant association between performance on the CST's target absent trials and extraversion, $r(60) = -.26$, $p = .038$, with this negative relationship being opposite to predictions based on previous research (Li et al., 2010). In addition, small to moderate associations were found between extraversion and both overall performance, $r(60) = -.18$, $p = .157$, and misses, $r(60) = -.18$, $p = .166$, on the CST, although these did not reach statistical significance. Note that reported *p*-values values are uncorrected and would fail to reach statistical significant if alpha levels were adjusted to account for running multiple correlations.

Although unrelated to our searching task, we considered the relationship between face matching and memory abilities for completeness. Previous research has suggested a moderate to large association between the GFMT and CFMT ($r_s = .39$ – Balsdon, Summersby, Kemp, & White, 2018; $r = .47$ – Robertson, Jenkins, & Burton, 2017; $r = .45$ to $.50$ – McCaffery et al., 2018). Here, analysing data from our whole sample rather than the unfamiliar judges only (given that no judges were familiar with the models in these two tests), we found a moderate correlation between the two measures of overall percentage correct, $r(94) = .34$, $p < .001$.

2.2.4 Familiarity and performance on the CST

The CST was designed to investigate the task of searching for an unfamiliar target in a chokepoint video. However, the nature of our data collection provided us with the opportunity to consider familiarity since a subsample of our judges was in the same student cohort as many of the targets. This familiarity with targets mirrors real-world searches in which, for example, officers may be selected for their previous experience with known troublemakers. Therefore, for those judges who reported some level of prior familiarity with at least one of the targets in their post-CST ratings (N = 35), we investigated whether familiarity resulted in increased response accuracy.

To illustrate, we considered trial-level accuracy for this subsample (averaging across judges) for each level of familiarity, finding that correct responses were more frequent as rated familiarity with the target increased (see Figure 2). Next, we analysed these data using a generalised linear mixed-effects model.

Following the same analysis strategy as for confidence (presented earlier) but simply replacing confidence with familiarity ratings, our fixed effects were the intercept, the effect of the familiarity rating, the effect of the trial type, and the familiarity x trial type interaction. Only the intercept in this model varied randomly across trials and judges, with a convergence failure when random slopes were included. The familiarity ratings given by judges were group mean centred.

Focussing on the fixed effects, we found that prior familiarity significantly predicted accuracy, $b = 0.25$, $SE = 0.08$, $p = .001$, with higher familiarity ratings associated with correct responses. The effect of trial type, $b = -0.27$, $SE = 0.36$, $p = .461$, and the familiarity x trial type interaction, $b = 0.10$, $SE = 0.11$, $p = .338$, were not statistically significant predictors.

2.2.5 Incidental learning during the CST

Due to how the chokepoint videos were created, many distractor individuals (those who made up the crowds) appeared in multiple videos. In addition, 15 targets also appeared in videos as distractors prior to the trial in which they served as a target. Although not ideal, this was unavoidable within the context of filming the same module's lectures over several weeks, a decision taken in order to best address difficulties with obtaining informed consent from all targets and distractors.

While perhaps unlikely, this meant that judges may have inadvertently learned to recognise targets from prior exposure, resulting in potentially increased accuracy on trials where the target had been seen previously as a distractor. To address this concern, we ran a modified version of the CST online, with each judge only completing one of the 20 trials (for details, see the Supporting Information). In this way, there could be no learning or recognition due to previous exposure. Our results suggested that relative trial difficulties were not affected by prior target exposure on some trials (comparison of 'full test' and 'one trial' judges found a correlation of .65 for trial-level accuracies). In addition, we found no evidence that 'full test' judges performed better overall or specifically on trials where the targets had previously been seen.

2.3 Discussion

Our results demonstrate that searching for unfamiliar faces in crowds at chokepoints is a very difficult task. As Table 1 summarises, overall performance (39%) was poor. In fact, on both target-present and target-absent trials, judges were far more likely to make an error than a correct decision. Further, on target-present trials, judges were more likely to choose a distractor rather than the target themselves. These levels of performance were notably lower than the hits (0.67) and correct rejections (0.52) reported by Mileva and Burton (2019), for example. This highlights the contrast between CCTV footage review (with pause and rewind enabled) and searching in real time.

We found that trial-by-trial confidence ratings predicted response accuracies, suggesting that judges had some insight into their abilities. This result mirrors previous work with face matching tasks (Bruce et al., 1999; Hopkins & Lyle, 2019). Interestingly, we found a stronger relationship for target-present trials, perhaps illustrating that judges found it easier to determine their confidence in a selection once an individual had been chosen. When responding 'absent', it may be more difficult to quantify one's certainty in relation to the specific crowd video just seen.

In line with the results of Davis and colleagues (2018) with the SFCT, we found evidence of a moderate association between CST performance and face memory ability. In addition, our results demonstrated an overlap (although again, moderate) between search performance and face matching. Taken together, relationships between face searching performance and tests of face matching and memory are present but limited, suggesting additional underlying mechanisms involved in searching for faces.

We also found small to moderate associations between extraversion and CST performance, although these were in the opposite direction to those predicted by previous research (Li et al., 2010). No associations were found with the remaining four personality facets, despite previous evidence that neuroticism may influence face processing tasks (Megreya & Bindemann, 2013). Of most relevance to the current work, Davis and colleagues (2018) found no relationship between personality factors and performance on their SFCT task. As such, our contradictory findings with extraversion require replication before any conclusions should be drawn.

Under certain conditions, law enforcement and other professionals may be faced with searching for familiar targets, and here, we had the opportunity to consider how familiarity might influence performance. Previous research has shown that familiarity with the target resulted in substantial improvements in face matching when comparing photographs with CCTV still images (Bruce, Henderson, Newman, & Burton, 2001). Indeed, this is why many travelling football teams in the UK, for instance, are accompanied by police or security personnel who are familiar with known troublemakers within the team's particular fan base. Our results replicated this effect for

searching, with increased prior familiarity with the target resulting in increased likelihood of a correct response.

3 Experiment 2 – Improving performance through exposure to variability

Previous research suggests that providing variability information, through multiple images of an unfamiliar face, may improve both matching (White, Burton, Jenkins, & Kemp, 2014; cf. Ritchie et al., 2019) and learning (Ritchie & Burton, 2017) performance, with larger variation across images of the same individual producing increased performance (Menon, White, & Kemp, 2015; Ritchie & Burton, 2017). Further, initial evidence suggests that viewing multiple images of a target may also prove beneficial when searching for that person in a crowd (Mileva & Burton, 2019). Here, we presented multiple images for each target, considering both low (images taken within minutes of each other) and high variability images (social media images that were unconstrained) in order to determine whether the amount of variability would affect performance.

3.1 Method

3.1.1 Participants

Fifty-one university students and other volunteers (33 women; age M = 24.65 years, SD = 11.23; 96% self-reported ethnicity as White) took part in exchange for course credits or monetary compensation. The data from 42 additional judges were excluded as they reported familiarity with at least one of the targets.

All judges provided written informed consent before taking part, and received both written and verbal debriefings at the end of the experiment. There was no overlap between this sample and those who took part in Experiment 1.

3.1.2 Stimuli

During target photography, we collected social media images of our targets in addition to filming a short conversation. The three low variability images comprised still frames taken at equal intervals from the conversation videos, whereas the three high variability images were the first three social media images submitted by the target that displayed the face clearly and unobscured, and at an acceptable resolution. All images were cropped to show only the head and shoulders (see Figure 3).

The video stimuli were those used in Experiment 1.

3.1.3 Procedure

Other than the presentation of three images of the target during each trial rather than one, all details of the CST remained unchanged. Judges were assigned to condition (low variability: $N = 26$; high variability: $N = 25$) using an alternating system based on when they participated. No additional measures were included.

3.2 Results

In order to investigate whether CST performance for judges unfamiliar with the targets was improved with exposure to variability, one-way (Condition: single image, low variability, high variability) between-subjects analyses of variance (ANOVA) were carried out. The 'single image' data were provided by the 62 unfamiliar judges who carried out the original version of the chokepoint search test (see Table 1).

First, considering the percentage correct, we found significant differences across the three conditions, $F(2, 110) = 11.61$, $p < .001$, $\eta^2_p = 0.17$. Pairwise comparisons (Bonferroni corrected

here and below) revealed that low variability images produced an improvement over both single image ($p < .001$) and high variability conditions ($p = .001$). However, these latter two conditions did not differ from each other ($p = 1.00$). Table 3 summarises these results.

Next, we carried out the same analyses for the five types of responses given by judges. For hits, we found the same pattern of results as for percentage correct, with a significant difference across conditions, $F(2, 110) = 5.94$, $p = .004$, $\eta^2_p = 0.10$, and pairwise comparisons revealing that low variability produced a greater proportion of hits than the other two conditions (both ps $< .009$), which did not differ from each other ($p = 1.00$). For misidentifications, we found an effect of condition, $F(2, 110) = 4.21$, $p = .017$, $\eta^2_p = 0.07$, with low variability producing a lower proportion of misidentifications than for the single image condition ($p = .015$). The remaining comparisons were not significant (both ps $> .134$). Finally, for misses, we found no difference across conditions, $F(2, 110) = 0.64$, $p = .527$, $\eta^2_p = 0.01$.

For correct rejections, we again found a significant difference across conditions, $F(2, 110) = 9.06$, $p < .001$, $\eta^2_p = 0.14$. Pairwise comparisons revealed that low variability produced a greater proportion of correct rejections than the single image ($p < .001$) and high variability conditions ($p = .026$). The latter two conditions did not differ from each other ($p = .908$). Given that target-absent trial responses must either be correct rejections or false positives, analyses of both response types produced identical results.

3.3 Discussion

It is unclear whether providing multiple images during face matching produces measurable benefits (White, Burton, et al., 2014; Ritchie et al., 2019), although evidence suggests that it improves searching for targets in crowds (Mileva & Burton, 2019). Here, we found that three low (but not high) variability images improved face searching performance over a single image.

Although recent work proposes that higher variability should be expected to produce greater benefits (Menon et al., 2015), we suggest that our result is likely due to recency and context rather than variability. As Figure 3 illustrates, the high variability images clearly provided more information regarding the different ways in which the target might appear. However, the low variability images (still frames from a short video) depicted the target as they appeared around the time that the crowd videos were filmed – all crowd and target materials were collected within the same semester. In addition, these images were taken from a video that was filmed on campus, perhaps better matching the context in which the crowd videos were filmed with regard to hairstyle, make up, clothing, etc. Social media photographs, in contrast, may be less likely to mirror the target's appearance during lectures.

We therefore interpret these results as evidence that performance can be improved through providing information about within-person variability. However, it is clear that there are situations in which more variability does not result in better performance. That the social media images were more varied, but less recent, than the video still frames means that we cannot draw any firm conclusions regarding the lack of a benefit. Therefore, we invite further exploration into the potential effects of variability versus recency.

4 General discussion

We aimed to simulate the task of searching for a person in a crowd in real time, a scenario commonly faced by security personnel at live events and in other situations. While recent research has started to consider how people search for faces through the review of CCTV footage (Davis et al., 2018; Mileva & Burton, 2019), this task of real-time search has received no attention to date.

Our focus here was on searching with the benefit of a natural chokepoint – a doorway. When available, such contexts allow for a more controlled, and perhaps thorough, search for the target since all individuals pass by in approximately single file. As such, this context may differ from

more unconstrained crowd searches, where people are able to walk in any direction and several individuals can pass by at once. For unconstrained crowds, it seems likely that other types of information will play a role in identification. With increased viewing distance, for example, evidence suggests that both the face and body independently contribute to recognition accuracy, although at closer distances, people rely only on the face (Hahn, O'Toole, & Phillips, 2016). While our videos always provided relatively clear views of faces, it would be interesting to explore other situations in which individuals can also be seen as they approach the chokepoint, e.g., when tourists filter through metal detectors that are placed in open spaces.

The results of Experiment 1 provide an initial investigation into the difficulties faced by judges, demonstrating that performance levels were poor and the majority of responses were incorrect. In particular, it was concerning to find that false positives were more common than correct rejections when targets were absent, having potential implications for real-world performance. Of course, additional factors almost certainly play a role in real chokepoint searches, including the motivation to spot a specific target. However, attempts to alter motivation, at least within the laboratory (presenting the target as "wanted" versus "missing persons"), have failed to influence levels of performance (Mileva & Burton, 2019).

Interestingly, both face matching and memory were only moderately predictive of searching performance, while any associations with personality facets remained unclear or absent. Previous work has suggested that performance on a test of face matching may be unable to predict performance on body- and biological motion-based tests, perhaps highlighting the limitations of focussing on face abilities when attempting to determine how people may perform with other types of identification (Noyes, Hill, & O'Toole, 2018). Researchers might also consider whether other individual differences might better predict search abilities, and related, whether those who are established as high performers on other face tasks (e.g., super recognisers; Russell et al., 2009) will also excel on this type of test.

Our results demonstrated that trial confidence predicted subsequent accuracy, mirroring work with face matching (Bruce et al., 1999; Hopkins & Lyle, 2019). Given that self-reports of general face processing skills are only moderately related to actual performance (e.g., Bobak et al., 2016; Palermo et al., 2017), future research might consider investigating self-report measures of how well people believe they will perform on face searching tasks prior to completing such tests.

We also found that increased prior familiarity with a target resulted in significant improvements in accuracy. This finding illustrates the nature of familiarity as a continuum (Kramer, Young, et al., 2018), as well as emphasising the recent focus in this field on the distinction between familiar versus unfamiliar face processing (Burton, 2013). Indeed, early work in this area has shown that those familiar with targets were able to identify them in even poor-quality video (Burton, Wilson, Cowan, & Bruce, 1999), highlighting just how powerful familiarity can be. These results also support the suggestion that target familiarisation as part of the experimental procedure can improve search performance (Davis et al., 2018), with future research needed to explore how best to familiarise viewers with a target beforehand.

In Experiment 2, we found that multiple low, but not high, variability images produced improvements in searching in comparison with single images. Although research suggests that higher variability should be expected to produce greater benefits (Menon et al., 2015; Ritchie & Burton, 2017), the current result is likely due to recency and/or context rather than variability. While multiple images may result in benefits in terms of the information provided of the target's face, there are clearly situations in which larger variability does not necessarily produce better performance (e.g., when that variability informs regarding how someone looked several years ago). This caveat to the findings of previous work is something that should be investigated further as it speaks to the selection of images when carrying out the search for a suspect or victim in real-world cases.

Along with providing multiple instances of a target in order to improve search accuracy (Mileva & Burton, 2019), there is reason to suggest that an average image of the target (a computer

blend of several instances) may also aid searches, at least in static arrays of images (Dunn, Kemp, & White, 2018). These averages are thought to provide a stable face representation by diluting idiosyncratic aspects of particular instances (Jenkins & Burton, 2011), therefore providing greater identity information. Although evidence from face matching tasks is mixed with regard to whether averages improve performance (Ritchie et al., 2018, 2019; White, Burton, et al., 2014), the potential benefits of their use during crowd searching has yet to be investigated.

We acknowledge that the majority of distractors within our crowd videos were White women, and for this reason, we selected only targets from this demographic. In order to explore own-ethnicity face searching, we recruited mostly White judges, with the assumption based on work with face recognition (Meissner & Brigham, 2001) and matching (Megreya, White, & Burton, 2011) that other-ethnicity judges would likely demonstrate lower levels of accuracy on our test. However, further work is required in order to confirm this assumption. Similarly, our judges were undergraduate students for the most part, and with face learning abilities peaking in the 30s (Germine, Duchaine, & Nakayama, 2011), it remains unclear how judges in other age groups may perform.

The apparent difficulties highlighted in the current work by the low performance measured during the CST provide the opportunity for training and improvement. Perhaps certain training programs may find some success with increasing performance on this type of test, although there has been little progress in this area to date with regard to training and face matching (Towler et al., 2019). One promising start, as noted earlier, might be to consider the process of familiarisation with the target through examination and discussion regarding the images prior to search, as demonstrated by recent work (Davis et al., 2018). Perhaps another potential route to improved performance can be found through working in pairs, which has already been shown to be beneficial when performing face matching tasks (Dowsett & Burton, 2015).

In conclusion, the current work represents the first investigation into the challenge of searching for a face in a crowd in real time. While commonly encountered by security personnel in

several professions, no prior research has considered this task, and our results highlight its difficulty. We hope that our work motivates further research into this important area of person identification.

Data availability statement

The data that support the findings of these experiments are available from the corresponding author upon reasonable request.

References

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications, 3*, 25.

Bate, S., & Dudfield, G. (2019). Subjective assessment for super recognition: An evaluation of self-report methods in civilian and police participants. *PeerJ, 7*, e6330.

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: Principles and Implications, 3*, 22.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bindemann, M., Avetisyan, M., & Blackwell, K. -A. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*(4), 378-386.

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology, 72*(4), 872-881.

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.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66*(8), 1467-1485.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202-223.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286-291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243-248.

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*(4), 339-360.

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*(3), 207-218.

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

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*(2), 192-203.

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology, 106*(3), 433-445.

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*(4), 576-585.

Dunn, J. D., Kemp, R. I., & White, D. (2018). Search templates that incorporate within-face variation improve visual search for faces. *Cognitive Research: Principles and Implications, 3*, 37.

Durova, M. L., Dimou, A., Litos, G., Daras, P., & Davis, J. P. (2017, December). *TooManyEyes: Super-recogniser directed identification of target individuals on CCTV*. Paper presented at the 8th International Conference on Imaging for Crime Detection and Prevention, Madrid, Spain (pp. 43-48). Stevenage, UK: IET.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118*(2), 201-210.

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

Gray, K. L., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science, 4*(3), 160923.

Hahn, C. A., O'Toole, A. J., & Phillips, P. J. (2016). Dissecting the time course of person recognition in natural viewing environments. *British Journal of Psychology, 107*, 117-134.

Hopkins, R. F., & Lyle, K. B. (2019). Image-size disparity reduces difference detection in face matching. *Applied Cognitive Psychology*. Advance online publication.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1571), 1671-1683.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*(3), 211-222.

Kramer, R. S. S., Mireku, M. O., Flack, T. R., & Ritchie, K. L. (2019). Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications, 4*, 28.

Kramer, R. S. S., Mohamed, S., & Hardy, S. C. (2019). Unfamiliar face matching with driving licence and passport photographs. *Perception, 48*(2), 175-184.

Kramer, R. S. S., Mulgrew, J., & Reynolds, M. G. (2018). Unfamiliar face matching with photographs of infants and children. *PeerJ, 6*, e5010.

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition, 172*, 46-58.

Lander, K., & Poyarekar, S. (2015). Famous face recognition, face matching, and extraversion. *Quarterly Journal of Experimental Psychology, 68*(9), 1769-1776.

Li, J., Tian, M., Fang, H., Xu, M., Li, H., & Liu, J. (2010). Extraversion predicts individual differences in face recognition. *Communicative & Integrative Biology, 3*(4), 295-298.

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.

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology, 25*(1), 30-37.

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

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364-372.

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.

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology, 64*(8), 1473-1483.

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), 3-35.

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception, 44*(11), 1332-1341.

Mileva, M., & Burton, A. M. (2019). Face search in CCTV surveillance. *Cognitive Research: Principles and Implications, 4*, 37.

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: Using individual data in the interpretation of group results. *Cognitive Research: Principles and Implications, 3*, 23.

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., ... & McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology*, *70*(2), 218-233.

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics, 76*(5), 1335-1349.

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http: //www.R-project.org/

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 897-905.

Ritchie, K. L., Mireku, M. O., & Kramer, R. S. S. (2019). Face averages and multiple images in a live matching task. *British Journal of Psychology*. Advance online publication.

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141*, 161–169.

Ritchie, K. L., White, D., Kramer, R. S. S., Noyes, E., Jenkins, R., & Burton, A. M. (2018). Enhancing CCTV: Averages improve face identification from poor-quality images. *Applied Cognitive Psychology, 32*(6), 671-680.

Robertson, D. J., Jenkins, R., & Burton, A. M. (2017). Face detection dissociates from face identification. *Visual Cognition*, *25*(7-8), 740-748.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252-257.

Shah, P., Sowden, S., Gaule, A., Catmur, C., & Bird, G. (2015). The 20 item prosopagnosia index (PI20): Relationship with the Glasgow face-matching test. *Royal Society Open Science, 2*(11), 150305.

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE, 14*(2), e0211037.

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

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20*(2), 166-173.

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

Table 1. Performance on the CST, along with tests of face matching and memory, for unfamiliar judges in Experiment 1

Task	Measure	*M*	95% CI
CST	Percentage correct	38.71	[34.90, 42.52]
	TP: Hits	0.34	[0.30, 0.38]
	TP: Misidentifications	0.44	[0.39, 0.49]
	TP: Misses	0.22	[0.18, 0.27]
	TA: Correct rejections	0.43	[0.38, 0.49]
	TA: False positives	0.57	[0.51, 0.62]
GFMT	Percentage correct	84.07	[81.76, 86.38]
	Sensitivity index, *d'*	2.26	[2.06, 2.45]
	Criterion, *c*	0.01	[-0.10, 0.11]
CFMT	Percentage correct	73.54	[70.44, 76.65]

Note. CST = Chokepoint Search Test; GFMT = Glasgow Face Matching Test; CFMT = Cambridge Face Memory Test; TP = target present; TA = target absent.

Table 2. Correlations between CST performance and other measures for unfamiliar judges in Experiment 1

			1	2	3	4	5	6
1	CST	Percentage correct	–					
2		TP: Hits	**.68** [.51, .79]	–				
3		TP: Misidentifications	**-.85** [-.91, -.76]	**-.54** [-.70, -.34]	–			
4		TP: Misses	**.37** [.13, .57]	**-.27** [-.49, -.02]	**-.66** [-.78, -.49]	–		
5		TA: Correct rejections	**.85** [.76, .91]	.19 [-.06, .42]	**-.74** [-.84, -.60]	**.68** [.52, .80]	–	
6		TA: False positives	**-.85** [-.91, -.76]	-.19 [-.42, .06]	**.74** [.60, .84]	**-.68** [-.80, -.52]	**-1.00**	–
	GFMT	Percentage correct	.22 [-.03, .44]	**.29** [.04, .50]	-.17 [-.40, .08]	-.06 [-.31, .19]	.09 [-.16, .33]	-.09 [-.33, .16]
		Sensitivity index, *d'*	.20 [-.06, .43]	**.27** [.02, .49]	-.16 [-.40, .09]	-.06 [-.30, .20]	.07 [-.18, .31]	-.07 [-.31, .18]
		Criterion, *c*	-.05 [-.29, .21]	-.04 [-.28, .22]	-.03 [-.27, .22]	.06 [-.19, .31]	-.04 [-.28, .22]	.04 [-.22, .28]
	CFMT	Percentage correct	.20 [-.06, .43]	**.26** [.01, .48]	-.14 [-.38, .11]	-.07 [-.31, .19]	.08 [-.18, .32]	-.08 [-.32, .18]
	Mini-IPIP	Extraversion	-.18 [-.41, .07]	.03 [-.22, .28]	.13 [-.12, .37]	-.18 [-.41, .08]	**-.26** [-.48, -.02]	**.26** [.02, .48]
		Agreeableness	-.08 [-.32, .17]	-.05 [-.30, .20]	.06 [-.20, .30]	-.02 [-.27, .23]	-.07 [-.32, .18]	.07 [-.18, .32]
		Conscientiousness	.00 [-.25, .25]	.03 [-.22, .28]	-.09 [-.33, .16]	.08 [-.18, .32]	-.03 [-.27, .23]	.03 [-.23, .27]
		Neuroticism	-.05 [-.29, .20]	-.01 [-.26, .24]	.13 [-.12, .37]	-.15 [-.38, .11]	-.06 [-.30, .20]	.06 [-.20, .30]
		Intellect/imagination	.05 [-.20, .30]	-.09 [-.33, .17]	.07 [-.19, .31]	.00 [-.25, .25]	.13 [-.12, .37]	-.13 [-.37, .12]

Note. CST = Chokepoint Search Test; GFMT = Glasgow Face Matching Test; CFMT = Cambridge Face Memory Test; IPIP = International Personality Item Pool; TP = target present; TA = target absent. Square brackets represent 95% confidence intervals. Values in bold are statistically significant at an uncorrected alpha level of .05.

Table 3. Performance on the CST for the three-image conditions in Experiment 2

Measure	Low Variability ($N = 26$)	High Variability ($N = 25$)
Percentage correct	56.92 [50.46, 63.39]	40.20 [31.76, 48.64]
TP: Hits	0.47 [0.39, 0.55]	0.31 [0.22, 0.40]
TP: Misidentifications	0.29 [0.20, 0.38]	0.42 [0.31, 0.53]
TP: Misses	0.24 [0.15, 0.34]	0.27 [0.20, 0.34]
TA: Correct rejections	0.67 [0.57, 0.77]	0.49 [0.39, 0.60]
TA: False positives	0.33 [0.23, 0.43]	0.51 [0.40, 0.61]

Note. Values appear as *M* [95% CI]. TP = target present; TA = target absent.

Figure Captions

Figure 1. Example trial from the chokepoint search test. The red box, drawn by the judge after freezing the video, illustrates who is thought to be the target. Here, the response is correct. [This target has given permission for her images to be reproduced here and in Figure 3.]

Figure 2. The effect of target familiarity on trial accuracy during the chokepoint search test.

Figure 3. Example images providing variability information. Either three low (top row) or high variability (bottom row) images were displayed during searches.