Low prevalence match and mismatch detection in simultaneous face matching: Influence of face recognition ability and feature focus guidance

Josh P. Davis, Callan Dray, Nikolay Petrov, and Elena Belanova

School of Human Sciences, Centre for Thinking & Learning, Institute of Lifecourse Development, University of Greenwich, London, UK, SE10 9LS

*Requests for reprints should be addressed to

Dr Josh P. Davis, School of Human Sciences, Face and Voice Recognition Lab, Institute of Lifecourse Development, University of Greenwich, London, UK, SE10 9LS

e-mail: j.p.davis@greenwich.ac.uk, https://orcid.org/0000-0003-0017-7159

This research was approved by the University of Greenwich Psychology Research Ethics Panel (14 May 2020).

Word count: 10308

Please note that this paper was accepted for publication in Attention, Perception & Psychophysics on 30 June 2021.

Abstract

Simultaneous face matching to verify identity is key to security and policing. However, matching is error-prone, particularly when target item prevalence is low. Two experiments examined whether superior face recognition ability and the use of internal or external facial feature guidance scales would reduce low prevalence effects. In Experiment 1, super-recognisers (n = 317) significantly outperformed typical-ability controls (n = 452), while internal feature guidance enhanced accuracy across all prevalence conditions. However, an unexpected effect in controls revealed higher accuracy in low prevalence conditions, probably because no low match or low mismatch prevalence information was provided. In Experiment 2, top-end-of-typical range ability participants (n = 841) were informed of their low prevalence condition and demonstrated the expected low prevalence effects. Findings and implications are discussed.

Keywords: Low prevalence effect; face matching; internal features, external features, super-recognisers

Introduction

Matching photos of unfamiliar faces to verify identity is key to many security and policing operations. For instance, to reduce illegal entry at border control, detecting rare fraudulent passports, or a correct *low prevalence mismatch* between holder and passport image, is the primary aim (e.g., Susa, Michael, Dessenberger, & Meissner, 2019; Tummon, Allen, & Bindemann, 2020). When surveilling events using computerised face recognition systems, operators may receive regular algorithm alerts suggesting a match between someone in the crowd and a suspect's image on a database. Here, the task is to disregard large numbers of false alarms identifying innocent bystanders (Fussey & Murray, 2019), in order to make a correct *low prevalence match*.

Identifying familiar people in photos tends to be highly accurate (e.g., Burton, Wilson, Cowan, & Bruce, 1999). However, in the scenarios described above, most staff will be unfamiliar with those depicted. Unfamiliar simultaneous face matching tasks of this type involving no memory are surprisingly error-prone (e.g., Bruce, Henderson, Greenwood, Hancock, Burton, & Miller, 1999; Dowsett & Burton, 2015; Fysh & Bindemann, 2018; for a review see Robertson et al., 2019). Error rates of at least 20% are common when the task is to view two high-quality close-up unfamiliar facial images and decide whether they depict the same person or not (Burton, White, & McNeill, 2010; Henderson, Bruce, & Burton, 2001). Matching photographs to unfamiliar individuals present in person is equally error prone (e.g., Davis & Valentine, 2009; White, Kemp, Jenkins, Matheson, & Burton, 2014). Employment in a role in which checking photo-ID documents may be key does not confer an advantage. Passport officers and task-naïve students generate similar rates of face matching errors (White et al., 2014; White, Dunn, Schmid, & Kemp, 2015).

To correctly identify that the same person is depicted in two images, observers must disregard inherent *within-person-variability* (i.e., changes in age, facial expressions, camera viewpoints) (Burton, 2013), that naturally differs from photo to photo, in order to detect reliable same-person cues inferring a *match*. To identify that two different people are depicted, they must isolate the *between-person-variability* (Jenkins, White, Van Montfort, & Burton, 2011), or properties that might be shared by more than one person within the two images to detect reliable cues signalling a *mismatch*. When faces are highly *familiar* these tasks are trivial as we have prior knowledge of how salient cues may vary in photos and still signal identity. However, an observer has no prior knowledge as to the extent to which within-person or between-person variability may vary in *unfamiliar* individuals. Not surprisingly then, an increase in variation in between-image properties (i.e., different facial viewpoints, age), is associated with an increase in unfamiliar face matching errors (see Davis & Valentine, 2015 for a review of face matching research).

Low Match-Mismatch Prevalence Effects

When target items are infrequent, as was described in the workplaces above, they are even more likely to be missed than if the prevalence of matched and mismatched items is equal (e.g., Moore & Johnston, 2013; Papesh & Goldinger, 2014; Wolfe, Horowitz, & Kenner, 2005; Wolfe, Horowitz, Van Wert, Kenner, Place, & Kibbi, 2007; Papesh, Heisick, & Warner, 2018; Susa, Michael, Dessenberger, & Meissner, 2019; although see Bindemann et al., 2010; Stephens et al., 2017 for research finding no effects). Some of the first research examined these effects in airport x-ray baggage screening when the identification of target items of concern (i.e., terrorist materials) will be extremely rare. Wolfe et al. (2005 see also Wolfe et al., 2007) revealed that effects were driven by criterion shifts, as when target

prevalence was low, participants employed a conservative, cautious decision-making strategy, reducing both hit rates of targets and false alarms of non-targets.

Similarly, Papesh and Goldinger (2014, see also Papesh, Heisick, & Warner, 2018; Weatherford, Erickson, Thomas, Walker, & Schein, 2020) found that face matching errors increased from 20% to 45% under increasing low prevalence conditions in designs replicating passport officers attempting to identify rare, mismatched items. Response times were also shorter for inaccurate than accurate mismatch items, suggesting reduced scrutiny. Interventions such as providing feedback on incorrect decisions, options to reconsider decisions, and directions to deliberate, had little impact on low prevalence face mismatch target detection. Similarly, experience in a job role in which photo identity verification is a regular component of the workload does not appear to protect against low prevalence effects in face matching. For instance, Weatherford, Roberson, and Erickson (2021) have recently shown that professional screeners such as security staff, bartenders, and other identity verifiers are also susceptible to low prevalence effects.

Weatherford et al. (2020) argue that in face matching paradigms, when mismatched items are known to be infrequent, an emphasis is placed on the search for the more common match cues, so as to induce a more liberal criterion (tendency to respond "same" or "match") when assessing perceived within-person variability in two images. Attention on between-person-variability cues is reduced, and searches may be ended early if the liberal match thresholds are met. Opposite effects are found if matched items are known to be infrequent. It could be expected therefore that if no information is provided as to matched-mismatched item prevalence in advance, attenuation of low prevalence item effects might be expected, as participants would be more likely to search with equal rigour for match and mismatch cues.

Individual Differences in Face Matching Ability

Recent research has revealed substantial largely inherited individual differences in unfamiliar face recognition ability in the population (Shakeshaft & Plomin, 2015), with the so-called 'super-recognisers' (SRs) occupying approximately the top 2% of this spectrum (e.g., Russell et al., 2009). As a group, SRs outperform typical-ability controls at face matching (Bobak, Hancock, & Bate, 2016), short term face memory (Bate et al., 2018), long-term face memory (Davis, Bretfelean, Belanova & Thompson, 2020), and CCTV type-review tasks (Davis, Forrest, Treml & Jansari, 2018). They maintain their advantage with faces of different ethnicities (e.g., Robertson, Black, Chamberlain, Megreya & Davis, 2020) and children (Bate, Bennetts, Murray, & Portch, 2020; Belanova, Davis & Thompson, 2018).

Not surprisingly, therefore, some police forces and identity verification businesses have successfully deployed SRs to roles in which identity verification is a key component of their daily activities (Davis et al., 2016; Davis et al., 2018; Robertson et al., 2016). SRs tend to make both more correct face match *and* mismatch decisions than controls (e.g., Bate, Frowd et al., 2018), suggesting that they possess superior skills at correctly assessing both within-person-variability and between-person-variability cues in two unfamiliar face images. Nevertheless, no published research appears to have investigated whether SRs' superiority reduces the impact of the low-prevalence bias in face matching.

Feature-Based Focus Guidance Interventions

Training interventions may also improve face matching performance. Professional forensic facial examiner training courses mainly advise trainees to break down the similarity of individual facial features in the images being compared, often with the aim of preparing a report for court (Megreya & Bindemann, 2018; Towler, White, & Kemp, 2017; see also White, Phillips, Hahn, Hill, & O'Toole, 2015). However, a forensic examiner, at work, will likely take many hours, or even days at face comparison, and these workplace conditions do

not match the fast decision-making employed by passport control officers or live face recognition algorithm operators.

Towler, White et al. (2017) also found that novice face matching accuracy was improved if participants first rated the similarity of 11 facial features on each pair of images. Results suggested that the ears were the most diagnostic feature. Other research isolating key features that might improve face matching accuracy has, however, reached different conclusions. Bruce et al. (1999) found higher accuracy when internal rather than external features were obscured. However, images were taken on the same day, allowing external cues such as hairstyle to be matched, which is less likely to be helpful with images separated by time. In contrast, Abudarham and Yovel (2016) identified lip thickness, hair colour, and eye colour; Megreya and Bindemann (2018), the eyebrows; and Zeinstra, Veldhuis, and Spreeuwers (2016), the chin/jawline as key features with high discriminative power. In addition, Kemp, Caon, Howard, and Brooks (2016) showed that masking external features improved face matching accuracy by 5% on difficult stimuli, suggesting that an internal feature focus strategy may be most beneficial.

Facial feature focus guidance might potentially also have a role in reducing the criterion shift impact seen in low prevalence target item conditions. If these shifts result from reduced scrutiny of salient low prevalence item cues as suggested by Weatherford et al. (2020), directing attention to the most discriminative facial features might have a positive impact. However, assessing 11 facial features before making judgments is time inefficient (see Towler et al., 2017). Therefore, the present research examined guided focus to three key internal or external features, selected as being some of the most discriminative in the studies described above. Unlike some research which has doctored image-parts to force attention to specific features (e.g., Bruce et al., 1999), full images remained visible throughout.

Current Research

In two experiments, the current research aimed to investigate whether a guided internal or external facial feature focus intervention would enhance face matching performance for low prevalence matched and mismatched trials in participants of different face recognition abilities. Participants were administered a 50-trial simultaneous face matching task and randomly allocated to different match-mismatch prevalence and feature-based guidance conditions. Not all face matching research has found low prevalence item effects (e.g., Bindemann et al., 2010), possibly because image sets were selected for their low within-person variability (i.e., photos were taken on the same day), which is unlikely to represent most identity verification workplace environments (Weatherford et al., 2020). Highly constrained carefully posed same-day image face matching may also require a different strategy to those commonly required in workplaces. Therefore, stimuli employed in the current research were mainly naturalistic, albeit close-up images taken in a range of different environments. Stimuli were also selected using pilot data to ensure overall equal difficulty for matched and mismatched trials in each prevalence and guidance condition. The aim was to ensure outcomes were not a consequence of some conditions containing harder trials than others.

In Experiment 1, no information as to 10-90%, 50-50%, 90-10% match-mismatch trial prevalence was provided to SRs and typical-range ability controls. In Experiment 2 the goal was to approximate real-life setting more closely, where it is likely that decision-makers will a) often be non-SRs and b) will roughly know the match-mismatch prevalence in the context they work in. Hence, participants were from the top end of the typical range at face recognition ability (no SRs), and correct information about matched and mismatched trial prevalence was provided.

In workplaces with low prevalence matched targets, ensuring high hit rates (identification of correct matched trials) will be most important, whereas achieving high correct rejection rates (CRs) (identification of correct mismatched trials) will be important in low prevalence mismatch target workplaces. These statistics were measured in the current research alongside signal detection theory measures of sensitivity (d') and criterion (C) (Green & Swets, 1966; Macmillan & Creelman, 1991), to measure discrimination of matched and mismatched trials, and to evaluate whether criterion shifts would replicate those found previously. In different research contexts, 'conservative' and 'liberal' criterion shifts may have different meanings. For clarity here, a conservative (positive) shift is one in which participants were cautious in responding 'same' and more likely to respond 'different'. A liberal bias was in the opposite direction. Analyses of response times (RTs) were also conducted. While the online survey system employed (Qualtrics) produces crude RT data in comparison to data collected in a lab, these data, collected from large numbers of participants, are nevertheless useful in examining the impact of using feature guidance scales.

Hypotheses were based on previous face matching research. First, SRs were expected to outperform controls (e.g., Russell et al., 2009). Second, performances were expected to be highest in the internal facial feature focus condition, followed by the external focus, and no guidance conditions respectively (e.g., Kemp et al., 2016). Third, compared to when matched-mismatched trial numbers were equal, responses of participants informed as to the low frequency of matched or mismatched items were predicted to be susceptible to the low prevalence criterion shift, so that they would be less likely to correctly respond to infrequent items (e.g., Weatherford et al., 2020).

Experiment 1

Experiment 1 examined the impact of varying match-to-mismatch prevalence and facial feature-focus guidance on the accuracy of SR's and typical-range-ability control's simultaneous face matching decisions. Group inclusion criteria were based on Cambridge Face Memory Test: Extended (CFMT+) (Russell et al., 2009) and Glasgow Face Matching Test (GFMT) (Burton et al., 2010) scores, and met criteria employed in previous research (e.g., Correll et al., 2020; Davis et al., 2020; Noyes et al., 2021; Satchell et al., 2019).

Hypotheses for face recognition ability and facial feature guidance matched those above, with SRs expected to outperform controls, while internal feature guidance was expected to have the strongest positive impact (Kemp et al., 2016). However, no information was provided to participants as to matched-mismatched low item prevalence. As such, we were agnostic as to the biasing impact on any criterion shift from this variable, even though it was likely that as the test progressed, participants, particularly SRs, would become increasingly aware of the relative prevalence of matched and mismatched items.

Method

Design

This research received ethical approval from the University of ANONYMIZED FOR REVIEW Psychology Research Ethics Panel. Participants completed 50 face matching trials in a 2 (Face recognition *Ability*: SRs *vs*. controls) x 3 (Mismatch *Prevalence*: 10%, 50%, 90%) x 3 (Facial feature-focus *Guidance* scales: External feature focus, internal feature focus, no guidance (control)) factorial design. SR and control group inclusion criteria were based on previously recorded CFMT+ (Russell et al., 2009) and GFMT scores (Burton et al., 2010); while allocation to prevalence and guidance conditions was random. Dependent variables were hit rates (proportion of correctly identified matched pairs), correct rejection rates (CRs, proportion of correctly identified mismatched pairs), and signal detection theory

(SDT) measures of sensitivity (d'), and criterion (C) or response bias (Green & Swets, 1966; Macmillan & Creelman, 1991). Response times (RTs) were collected. However, as RT data collection on Qualtrics is suboptimal and outlier prone, standardised z-scores were analysed and reported only in the supplementary materials.

Materials

Cambridge Face Memory Test: Extended (CFMT+) (Russell et al., 2009)*:* This 102-trial 4-block test has commonly been used to define super-recognition. In the first block, participants learn the faces of six white males displayed from different angles and with external features removed (i.e., hairstyle). Using a three-alternative choice design in the following two blocks, visual noise is included, and facial expressions vary. The fourth block adds repeating distractors and stronger visual noise. Scores out of 102 are produced.

Glasgow Face Matching Test (GFMT) (Burton et al., 2010): This 40-trial equal-match-mismatch prevalence simultaneous face matching test requires participants to decide whether male and female white faces are matched (20) or mismatched (20). There are virtually no memory demands. Scores out of 40 are produced.

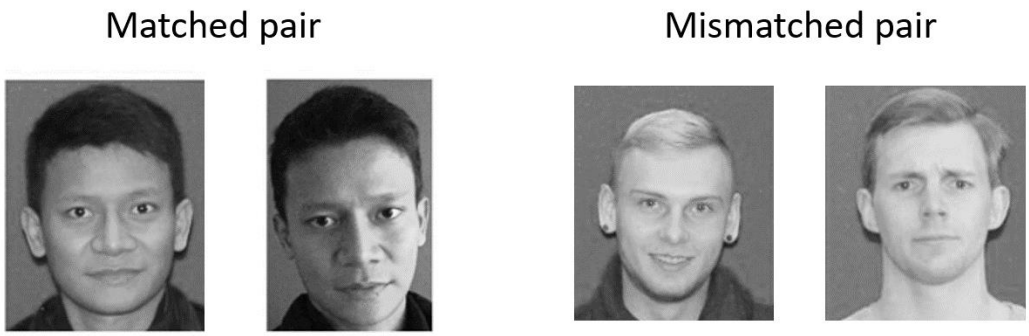
Simultaneous Face Matching Test (Varied Prevalence): The pairs of frontal facial stimuli for the three versions of this 50-trial test (10%, 50%, 90% mismatch) were selected from 233 trials (131 matched, 102 mismatched) employed in four unpublished tests created by the first author and taken by likely highly motivated individuals applying for identity verification jobs requiring superior face matching skills. All job applicants (hereafter pilot participants) (n = 248) had achieved scores of ≥ 38 on the GFMT as initial criteria for recruitment prior to taking the selection tests. This threshold is lower than those for the SR group in the current research (40 out of 40).

Face matching image sets possessing greater within- and between-person variance may be more likely to generate low prevalence item effects (Weatherford et al., 2020).

Therefore, the 233 pairs of male and female adult close-up images varied in age (from same day to up to 10 years apart), camera lens type, original distance from camera, facial angle, facial expressions, clothing, the wearing of paraphernalia or not (i.e., hats, glasses, jewellery). More than ten distinct self-defined ethnicities were represented (e.g., Chinese, Black, Indian), although depending on condition, slightly more than half were White-Caucasian (See Figure 1). Some images were slightly cropped obscuring tops of heads or edge of ears.

Figure 1.

Examples of matched and mismatched pairs from the Simultaneous Face Matching Test



Item analyses conducted on the pilot participant data ensured all individual items used in the final test had been identified at better than chance rates (> 0.50). Images were also selected to ensure virtually identical difficulty for matched and mismatched sets in each prevalence condition (mean hits and CRs in each condition = 0.88; range = 0.62 to 1.00 for individual image pairs); whilst wherever possible, the same image-pairs appeared (see Table 1). For instance, 5 mismatched pairs and 5 matched pairs appeared in all conditions. Twenty pairs were unique to the 10% mismatch, and 20 pairs to the 90% mismatch versions. The remaining 20 mismatched and 20 matched trials appeared in two versions only.

Table 1.

Distribution and mean pilot participant (n = 248) accuracy rates for matched (hits) and mismatched (CRs) identity pairs within each prevalence condition

	10% Mismatch	50% Mismatch	90% Mismatch
Matched pairs			
• Unique to Matched	20	-	-
• Matched and equal prevalence	20	20	-
• Displayed in all conditions	5	5	5
Pre-research item mean hits (*SD*)	0.88 (.08)	0.88 (.08)	0.88 (.08)
• Displayed in all conditions	5	5	5
• Mismatched and equal prevalence	-	20	20
• Unique to Mismatched	-	-	20
Pre-research item mean CRs (*SD*)	0.88 (.09)	0.88 (.09)	0.88 (.09)
• Total trials	50	50	50

Participants

Participants meeting inclusion criteria, none of whom had applied for jobs above, were randomly selected for an invite from the ANONYMIZED FOR REVIEW volunteer participant database. They were e-mailed and incentivised by a random prize draw. Excluded participants were those, who entered incorrect personal codes - meaning previous CFMT+ and GFMT scores could not be accessed (n = 17) - those who scored below chance levels - suggesting low motivation (i.e., < 25/50) (n = 28) – and those with extremely long test completion times (> 40 min for no guidance conditions, and > 80 min for feature guidance conditions) (n = 19).

The final sample of 769 participants (males = 254, females = 511, other = 2, missing = 2; M_{age} = 39.1, SD = 12.1, missing = 6; White-Caucasian = 582, 2nd most common ethnicity (Hispanic or Latino) = 46) met criteria for SR and typical-ability controls, employed in previous research (e.g., Correll et al., 2020). As shown in Table 2, on their first attempts, SRs had previously achieved maximum scores on the GFMT (Burton et al., 2010), and in the top 2% of a representative UK sample (n = 254, M = 70.7, SD = 12.3) on the CFMT+ (Bobak,

Pampoulov et al., 2016). Typical-ability controls had scored within 1 SD of the mean of Burton et al.'s (2010: n = 192, M = 32.5, SD = 9.7), and Bobak, Pampoulov et al.'s (2016), GFMT and CFMT+ samples, respectively.

Table 2.

Group inclusion criteria and mean scores on the CFMT+ and GFMT

	SRs (n = 317)		Controls (n = 452)					
CFMT+ criterion	≥ 95		58-83					
GFMT criterion	40		28-36					
	M	*SD*	*M*	*SD*	*df*	*t*	*d*	*p*
CFMT+ (max 102)	97.20	1.85	73.82	7.25	532.40	65.61	4.42	<.001
GFMT (max 40)	40.00	0.00	34.06	1.94	451.00	65.23	4.33	<.001

Procedure

Participants were invited to take the tests using a Qualtrics link. After providing informed consent, they were given brief instructions as to research requirements and asked to enter a code allowing access to their previous CFMT+ and GFMT scores. No participant in Experiment 1 was informed in advance of the randomly allocated prevalence manipulations.

Participants were first informed that they would be presented with 50 unfamiliar face-pairs varying in difficulty and were asked to decide if these faces were of the same individual or not. Depending on randomly allocated guidance condition, participants then received a set of brief instructions which told them that they would be asked to focus on internal or external features, or if in the control group, they were given no feature instructions. For instance, prior to starting the external focus guidance condition, instructions stated that, "results from previous studies have identified that focussing on external features such as the ears, face shape and chin/jawline (internal guidance condition = eyebrows, nose, mouth) improved performance levels in face matching tasks. Therefore, we ask that before making a final decision you rate each of these features." Each then completed the 50 trials. On each trial,

presented on a different page, participants first rated, “the similarity of the features below within this facial pair”, using a Likert scale (1: Very Dissimilar to 5: Very Similar). Note that scale data were not analysed. A 'not applicable' option was also provided on each scale (i.e., for trials in which ears were partially covered by hair). Immediately below the scales, a final question asked whether the photos were of the same person or not. This required a ‘same’ or ‘different’ response.

In the no guidance (control) condition, no feature rating scales were provided, and participants completed the same/different question only. Accuracy of these decisions formed the dependent variables (i.e., hit rates: same decisions: identification of correct matched trials; CR rates: different decisions: identification of correct matched trials). There were no time limits. Upon completion of the experiment, a debrief was provided including feedback listing scores out of 50. Overall, feature guidance condition participants took approximately 30 min to finish. Those in the no guidance conditions took approximately 12 min.

Results

A minority of SRs (n = 31; 10.2%), but no controls, achieved maximum scores of 50 out of 50. Four three-way Type-III ANOVAs examined the effects of ability (SRs, controls), prevalence (10% mismatch, 50% mismatch, 90% mismatch) and feature guidance (internal, external, none) on hits, CRs, sensitivity (d'), and criterion (C). Type-III ANOVAs were chosen as interactions were expected (Field, 2017). Scores were not normally distributed, although ANOVA is robust to these violations with large samples (Blanca et al., 2017). Nevertheless, Type-II ANOVAs were also conducted, with no differences in conclusions. To protect against Type-I errors, which we consider more costly than Type-II errors (Armstrong,

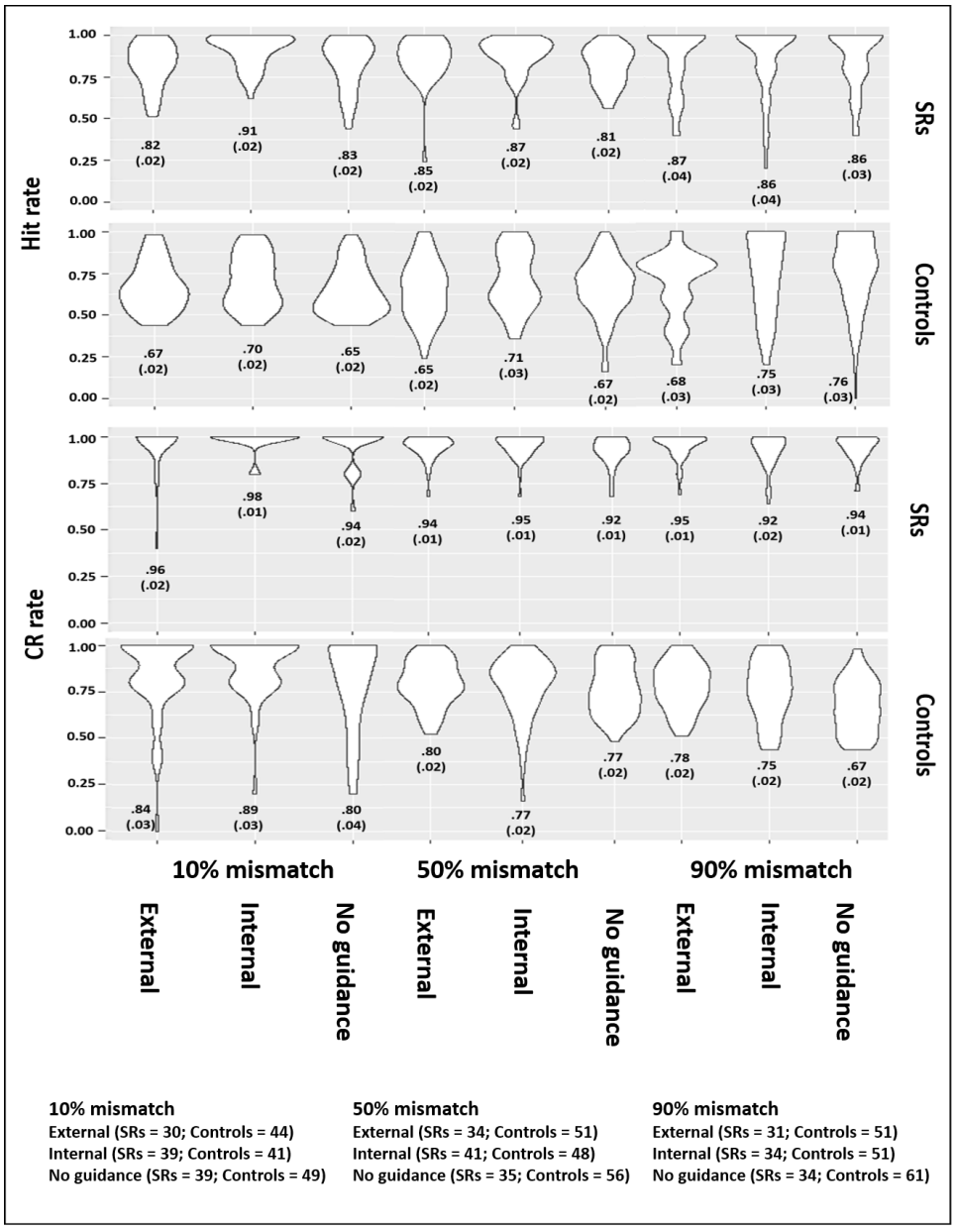
2014; Noble, 2009), post-hoc analyses employed the conservative Bonferroni correction to ensure retention of $\alpha = 0.05$.

Hits: This ANOVA revealed a significant main effect of ability, $F(1, 751) = 156.40$, $p < .001$, $\eta^2 = .172$. As expected, SRs ($M = .86$, $SD = .15$) outperformed controls ($M = .69$, $SD = .19$). The guidance main effect was significant, $F(2, 751) = 4.31$, $p = .014$, $\eta^2 = .011$. Paired comparisons found that as predicted, internal feature guidance generated higher hit rates than no guidance, $t(526) = 2.58$, $p < .05$, Cohen's $d = .21$, and external feature guidance, $t(493) = 3.10$, $p < .05$, Cohen's $d = .26$, while external feature guidance and no guidance conditions did not differ, $t(513) < 1$. There was a significant main effect of prevalence, $F(2, 751) = 3.01$, $p = .050$, $\eta^2 = .008$, and while post-hoc comparisons were not significant ($p > .05$), the trend in the data was opposite to that expected, as the highest hit rates were in the 90% mismatch (i.e. low match) condition. There were no significant interactions (all $F's < 1$).

CRs: This ANOVA revealed a significant main effect of ability, $F(1, 751) = 234.88$, $p < .001$, $\eta^2 = .238$. SRs ($M = .94$, $SD = .09$) outperformed controls ($M = .78$, $SD = .18$). The guidance main effect was significant, $F(2, 751) = 6.76$, $p = .001$, $\eta^2 = .018$. As expected, internal, $t(518) = 2.51$, $p < .05$, Cohen's $d = .35$, and external feature guidance, $t(512.08) = 3.07$, $p < .05$, Cohen's $d = .29$, generated higher CRs than no guidance, with no differences found for internal *vs*. external feature guidance, $t(493) < 1$. The prevalence main effect was significant, $F(2, 751) = 13.57$, $p < .001$, $\eta^2 = .035$. Outcomes were opposite to those predicted, as CRs were successively significantly higher in the 10% mismatch, than the 50% mismatch condition, $t(454.04) = 3.28$, $p < .05$, Cohen's $d = .31$, which were significantly higher than the 90% mismatch condition, $t(515.11) = 2.73$, $p < .05$, Cohen's $d = .20$.

Figure 2.

Mean Hits and CRs for SRs and controls in each condition (SEs in parentheses)



There was a significant ability x prevalence interaction, $F(2, 751) = 5.38$, $p = .005$, $\eta^2 = .014$. Two one-way ANOVAs revealed that controls, $F(2, 751) = 21.49$, $p < .05$, $\eta^2 = .054$, but not SRs, $F(2, 751) < 1$, generated significantly different CRs between prevalence conditions. As with the main effect described above, controls' CR rates were successively

significantly highest in the 10%, than the 50%, $t(227.09) = 2.75$, $p < .05$, Cohen's $d = .32$; and 90% mismatch conditions, $t(316) = 2.89$, $p < .05$, Cohen's $d = .33$ respectively. Therefore, the paradoxical low prevalence effect was found in controls but not in SRs.

Sensitivity (d'): This ANOVA revealed a significant main effect of ability, $F(1, 751) = 602.04$, $p < .001$, $\eta^2 = .445$. SRs ($M = 3.17$, $SD = .84$) outperformed controls ($M = 1.64$, $SD = .89$). There was a significant guidance main effect, $F(2, 751) = 14.03$, $p < .001$, $\eta^2 = .036$, whereby internal feature guidance generated higher sensitivity than no guidance, $t(526) = 4.71$, $p < .05$, Cohen's $d = .41$, and external feature guidance conditions, $t(493) = 3.19$, $p < .05$, Cohen's $d = .28$, which did not differ ($p > .05$). There was a significant prevalence main effect, $F(2, 751) = 19.86$, $p < .001$, $\eta^2 = .050$, with the 10% mismatch condition generating higher sensitivity than the 90%, $t(502) = 3.09$, $p < .05$, Cohen's $d = .28$, and 50% mismatch conditions, $t(505) = 5.33$, $p < .05$, Cohen's $d = .48$, which did not differ ($p > .05$).

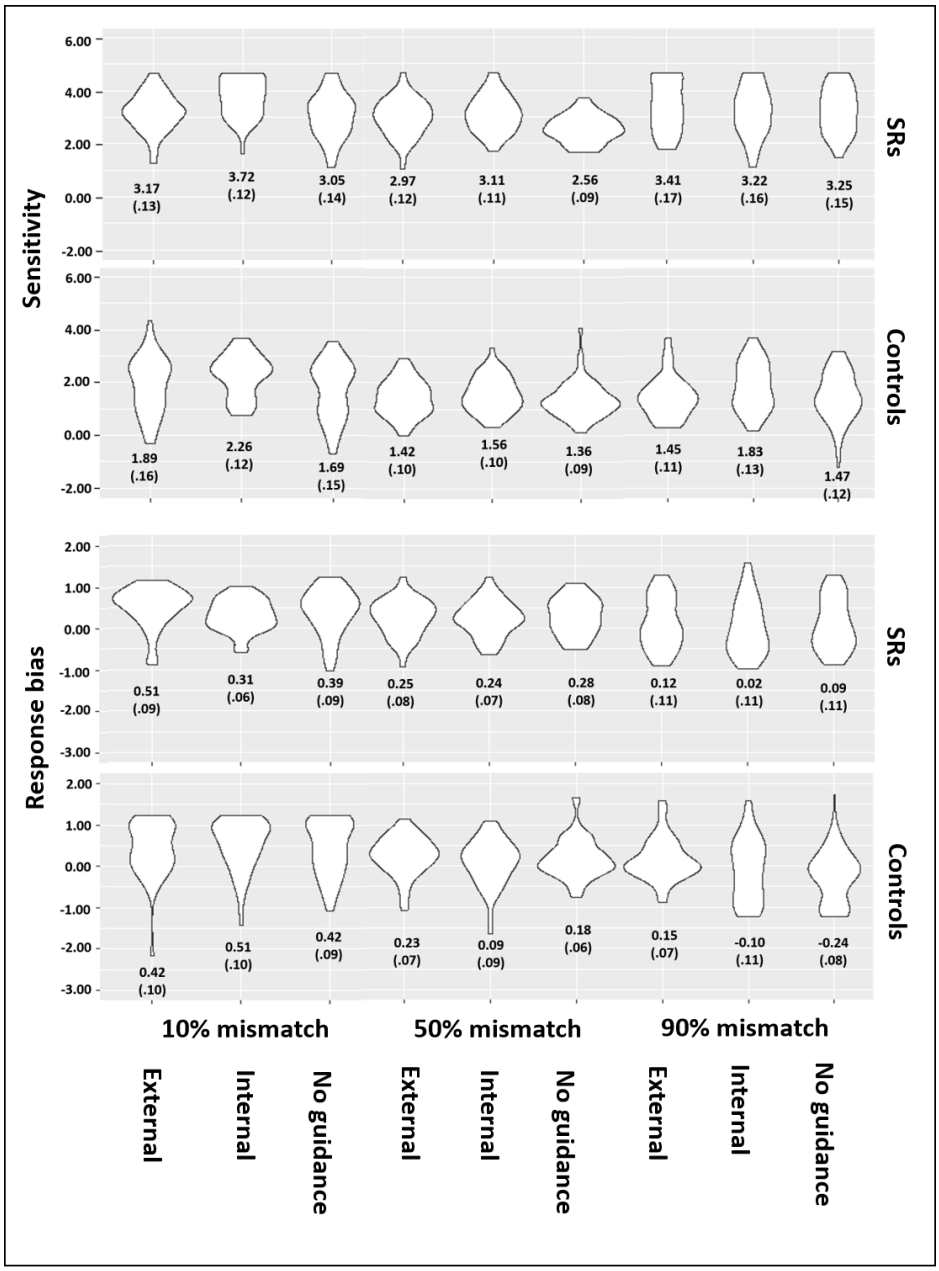
There were no significant interactions ($F's \leq 2.93$, $p > .05$).

Criterion (C): This ANOVA revealed only a significant prevalence main effect, $F(2, 751) = 32.12$, $p < .001$, $\eta^2 = .079$, so that successively, the 10% mismatch condition generated higher positive values of C (more likely to respond 'different') than the 50% mismatch condition, $t(471.79) = 4.59$, $p < .05$, Cohen's $d = .41$, while the 50% mismatch condition was more positive than the 90% mismatch condition, $t(480.09) = 4.41$, $p < .05$, Cohen's $d = .38$. There was no main effects of ability or guidance, $F's < 2.35$. These results suggest that performances were driven by a criterion shift in the opposite direction to predictions.

There were no other significant effects or interactions ($F's \leq 2.35$, $p > .05$).

Figure 3.

Mean sensitivity (d') and response bias (C) across all conditions (SEs in parentheses)



Discussion

Experiment 1 examined the impact of face recognition ability and external (ears, face shape, jawline) or internal (eyebrows, nose, mouth) feature focus guidance on low match-

mismatch trial prevalence effects in simultaneous face matching. As expected, with strong effect sizes, on all accuracy measures (hits, CRs, d'), as a group, SRs significantly outperformed controls, regardless of matched and mismatched trial prevalence. There were no between-group criterion effects. In addition, hits and sensitivity (d'), were significantly higher in the internal focus guidance condition, than in the control and external focus conditions, which did not differ. In contrast, CRs were roughly equal in the internal and external feature focus guidance conditions, with both significantly higher than in the no guidance conditions. There was no significant impact of guidance on criterion (C), or any interactions between guidance and other variables on any measure. These results provide support for the use of such scales over no scales, albeit effects were smaller than those comparing SRs and controls. Policy makers would need to decide whether the slight gains in accuracy offset the increased staff time to use such scales (≈ 30 min with scales *vs*. 12 min for no scales over 50 trials),

Unlike most previous research in this area, no warnings as to the low prevalence of items were provided to participants, albeit as the test progressed it might be expected that some high scoring participants, particularly SRs might have become aware of this imbalance. However, lower scoring participants would have been unlikely to have identified this manipulation. Analysis of criterion scores revealed a criterion shift in the opposite direction to previous low prevalence effects research in which such a warning has been provided (e.g., Papesh & Goldinger, 2014). Compared to the equal matched-mismatched control condition (50% mismatch); when mismatched trials were infrequent (10% mismatch), participants were significantly more likely to display a conservative cautious response bias (i.e. more likely to respond "different"). Similarly, when matched trials were infrequent (90% mismatch), participant's more liberal response bias suggested they were significantly more likely to respond 'same'.

Significant criterion shifts do not necessarily imply differences in accuracy rates for infrequent prevalence items. For this, inspections of hits and CRs are required. The prevalence main effect was significant with hits, albeit effect sizes were small, and no post hoc analyses were significant. Nevertheless, the non-significant trend in the data was consistent with the paradoxical criterion effects described above, and in the opposite direction to those that would be predicted based on previous low prevalence effects research. Hit rates tended to be slightly highest when infrequent in the 90% mismatched condition; while hit rates were lowest when relatively common in the 10% mismatched condition.

In terms of CRs, a significant interaction between prevalence and ability was revealed. With SRs, there were no significant effects of prevalence on CR rates. In controls, the results were consistent with the significant criterion shifts reported above. In contrast to predictions, when compared to the equal matched-mismatched condition, when mismatched trials were infrequent (10% mismatched), CRs were significantly higher. On the other hand, when mismatched trials were common (90% mismatched), CRs were significantly lower. Finally, sensitivity was also significantly highest in the 10% mismatch prevalence condition than the 90%, and 50% mismatch prevalence conditions, which did not differ.

Approximately 10% of SRs in Experiment 1 achieved maximum scores, while many others scored close to ceiling. This might explain the lack of significant effects on SR's CRs across the prevalence conditions, as well as the weak prevalence effects on hits in general. SRs are rare in the population, and organisations in which identity verification might be a key role might require large numbers of staff. Therefore, to reduce ceiling effects in Experiment 2, participants scoring in the 'top end of typical' at face recognition were invited. As such, participants who had generated scores in between those of Experiment 1's SRs and controls on the CFMT+ (i.e., 84-94) and GFMT (i.e., 37-39) were invited. The lower value on each test is approximately 1 SD above the normative means (Burton et al., 2010; Bobak et al.,

2016), expected to be achieved by about 16% of the population. In addition, most staff employed to perform identity verification tasks would almost certainly be aware of the likely prevalence of matched and mismatched Photo-ID. Therefore, participants in Experiment 2, were informed in advance as to the prevalence of each type of trial.

Experiment 2

Experiment 2 replicated and extended the design of Experiment 1. However, only participants achieving 'top end of typical' scores on the CFMT+ and GFMT were invited. Furthermore, unlike in Experiment 1, participants were informed of the prevalence of mismatched and matched items. The hypotheses were again based on low prevalence effects found in previous research (e.g., Papesh & Goldinger, 2014), and expected to contrast with the results of Experiment 1, as this time participants were explicitly aware of prevalence in advance. As such, a criterion shift was predicted to generate a more conservative response bias, with a tendency to respond "different" in low matched prevalence conditions (90% mismatch), reducing hits, and increasing CRs, whereas a liberal bias (tendency to respond "same") was expected in the low mismatched prevalence conditions (10% mismatch), increasing hits, and reducing CRs. Internal facial feature focus guidance was again expected to improve accuracy scores (hits and/or CRs) more than external guidance and no guidance, respectively.

Method

Design

The design of Experiment 2 was the same as Experiment 1 and employed a 3 (Mismatch *Prevalence*: 10%, 50%, 90%) x 3 (Facial feature-focus *Guidance* scales: external

feature focus, internal feature focus, no guidance) factorial design, with dependent variables: hit rates, CR rates, sensitivity (d'), response bias (C). Participants from Experiments 1 and 2 were also combined to compare performance across SRs, top-end of typical participants and typical range controls, with analyses reported in supplementary materials. Raw and standardised (z scores) response times (RTs) are also reported in supplementary materials.

Note, originally, Experiment 2 did not include the three 50% mismatch prevalence conditions and consisted of a 2 (Prevalence: 10%, 90%) x 3 (Guidance) design only. However, based on anonymous journal reviewer recommendations on an earlier version of this manuscript, we invited additional participants meeting Experiment 2's criteria, and randomly assigned them to the three 50% conditions, so as to better match Experiment 1's design. This change means that a fully randomised design cannot be reported, as there were two participant samples. However, results met expectations, and analyses on the participant samples suggest no differences on key criteria.

Participants

Participants were randomly selected from the same ANONYMIZED DATABASE as Experiment 1. All provided informed consent and permission to access previous scores on the CFMT+ and GFMT. Participants were excluded if they failed to provide participant codes (n = 6), scored below chance levels (< 25/50) (n = 9), or were extremely slow at completing the test (no guidance: > 36 min, feature guidance: > 91 min) (n = 35). The final sample comprised 841 participants (males = 285, females = 550, other = 3, missing = 3; M_{age} = 41.4, SD = 11.4, missing = 7; 668 White-Caucasian, 2nd most common ethnicity = 34 Hispanic or Latino). Participants CFMT+ scores ranged from 84 to 94 (M = 89.71; SD = 3.12) and GFMT scores ranged from 37 to 39 (M = 37.68; SD = 0.48).

Note: originally, we only recruited participants for the 10% (n = 272) and 90% (n = 295) Mismatch Prevalence conditions. Additional participants (n = 274) were subsequently recruited for the 50% Mismatch Prevalence condition. Independent t-tests showed that the two samples (i.e., those recruited to the 10% and 90% conditions *vs*. the participants allocated to the 50% Mismatch condition) reported similar ages (t's ≤ 1.27, p > .05), and similar scores on the CFMT+ (t's ≤ 1.48, p > .05), and GFMT (t's ≤ 1.95, p > .05). Chi squared tests also showed that both samples reported similar gender, χ^2 (1, n = 838) = 3.25, p > .05, and ethnicity proportions (White = 1, other = 0), χ^2 (1, n = 834) = 1.66, p > .05.

Materials and Procedure

The procedure was virtually identical to Experiment 1, except participants were correctly warned about the prevalence of matched and mismatched trials. Without giving exact proportions, at the start of the 10% mismatched condition, additional instructions informed participants that, "the majority of trials in this research are matched trials. In other words, most images depict the same person". Instructions at the start of the 50% mismatched condition informed that "the proportion of matched and mismatched trials are roughly equal. In other words, in approximately half of the trials both images depict the same person and in the other half of the trials, the two images depict two different people". Instructions at the start of the 90% mismatched condition informed that, **"**the majority of trials in this research are mismatched trials. In other words, most images depict different people."

Note: Following the additional participant recruitment for the 50% Mismatch trials, on analysing the data, the experimenters discovered that the feature guidance scales for the last two trials (Trial 49/50 out of 50) in the 50% Mismatch-External guidance condition of Experiment 1 had been wrongly labelled. All analyses reported in Experiment 1 were conducted with, and without, the inclusion of these two trials. There were no substantive

differences in outcomes on all analyses and as conclusions were identical, the data from all trials are reported.

Results

Fewer participants than in Experiment 1 achieved maximum scores of 50 out of 50 (n = 12; 1.4%). A similar Type-III prevalence (10% mismatch, 50% mismatch, 90% mismatch) x feature guidance (internal, external, none) ANOVA strategy to Experiment 1 was employed (with Type-II ANOVAs finding similar effects).

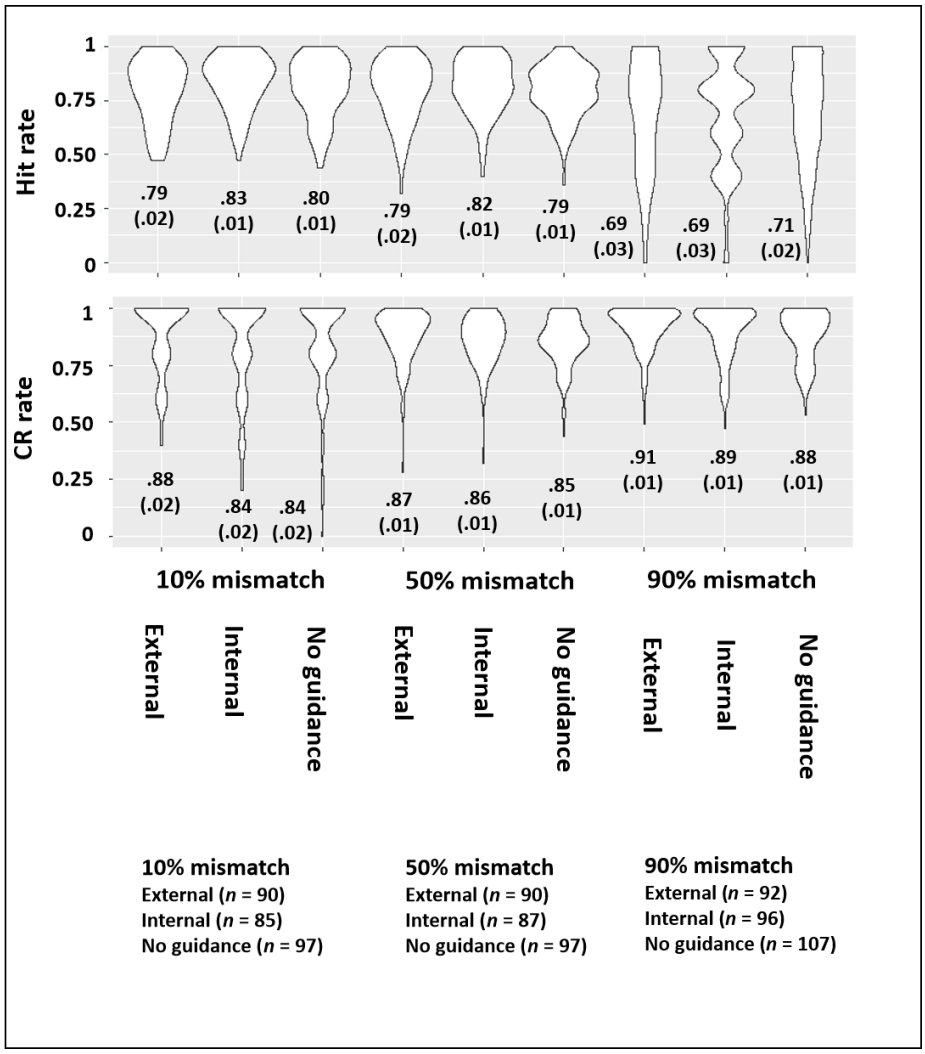
Hits: This ANOVA revealed a main effect of prevalence, $F(2, 832) = 8.65$, $p < .05$, $\eta^2 = .077$. In line with expectations, and counter to Experiment 1, hit rates were lower in the 90% mismatch condition ($M = .70$, $SD = .25$) compared to the 10% mismatch condition ($M = .81$, $SD = .14$), $t(463.06) = 6.60$, $p < .05$, Cohen's $d = .54$, and the 50% mismatch condition ($M = .80$, $SD = .13$), $t(449.21) = 6.16$, $p < .05$, Cohen's $d = .50$, whereas the latter two conditions generated similar performance, $t(544) < 1$. Other effects were non-significant $F(2, 561) < 1$.

CRs: This ANOVA revealed a main effect of prevalence, $F(2, 832) = 8.58$, $p < .05$, $\eta^2 = .014$. In line with expectations, and counter to Experiment 1, CR rates were lower in the 10% ($M = .85$, $SD = .20$), $t(414.148) = 2.81$, $p < .05$, Cohen's $d = .25$, and in the 50% mismatch condition ($M = .86$, $SD = .11$), $t(567) = 3.37$, $p < .05$, Cohen's $d = .27$, than in the 90% mismatch condition ($M = .89$, $SD = .11$). The 10% and 50% mismatch conditions generated similar CRs, $t(429.73) < 1$.

There was also a main effect of guidance, $F(2, 832) = 5.87$, $p < .05$, $\eta^2 = .009$. External feature guidance ($M = .89$, $SD = .13$) generated higher CRs than no guidance ($M = .86$, $SD = .15$), $t(571) = 2.68$, $p < .05$, Cohen's $d = .21$, while other comparisons were non-significant ($p > .05$).

Figure 4.

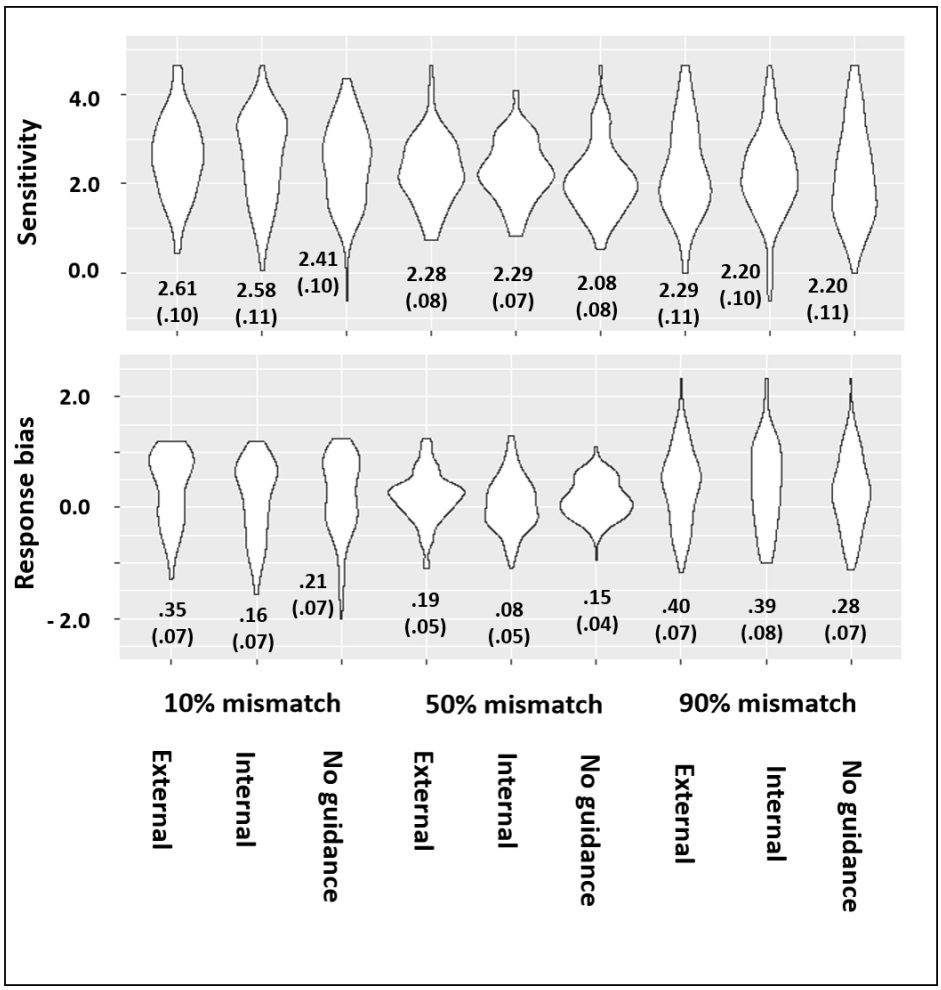
Mean proportions of Hit rates and CR rates across all conditions (SEs in parentheses)



Sensitivity (d'): This ANOVA revealed a main effect of prevalence, $F(2, 832) = 8.33$, $p < .05$, $\eta^2 = .025$. Sensitivity was greater in the 10% mismatch condition ($M = 2.53$, $SD = .94$), than in the 90% mismatch condition ($M = 2.23$, $SD = 1.03$), $t(565) = 3.63$, $p < .05$, Cohen's $d = .30$, and in the 50% mismatch condition ($M = 2.21$, $SD = .74$), $t(515.25) = 4.39$, $p < .05$, Cohen's $d = .38$, while the latter two conditions generated similar performance, $t(536.20) < 1$. Other effects were non-significant ($F's \leq 2.61$, $p > .05$).

Figure 5.

Mean sensitivity (d') and response bias (C) across all conditions (SEs in parentheses)



Criterion (C): This ANOVA revealed a main effect of prevalence, $F(2, 832) = 8.71$, $p < .05$, $\eta^2 = .021$. The 90% mismatch condition ($M = .35$, $SD = .71$) generated higher criterion scores (more likely to respond "different") than the 50% mismatch condition ($M = .14$, $SD = .43$), $t(489.18) = 4.34$, $p < .05$, Cohen's $d = .36$. Other comparisons were non-significant ($p > .05$), and there were no other significant effects ($F's \leq 2.56$, $p > .05$).

Comparison of the results from Experiment 1 and 2

Analyses were also conducted on the combined Experiment 1 and 2 data to compare the condition-impacted performances of the three ability groups. Three-way ANOVAs examined the effects of ability (Experiment 1 SRs: "SRs", Experiment 2 top-end-of-typical range participants: "Experiment 2 participants"; Experiment 1 typical-ability range "controls"), prevalence, and guidance on hits, CRs, sensitivity (d'), and criterion (C).

Hits: This ANOVA revealed a main effect of ability, $F(2, 1583) = 74.55$, $p < .001$, $\eta^2 = .086$. SRs outperformed controls, $t(751.36) = 12.90$, $p < .05$, Cohen's $d = .99$, and Experiment 2 participants, $t(698.78) = 8.04$, $p < .05$, Cohen's $d = .53$, while the latter also outperformed controls, $t(1291) = 6.59$, $p < .05$, Cohen's $d = .42$.

There was a main effect of guidance, $F(2, 1583) = 5.12$, $p = .006$, $\eta^2 = .006$. Internal feature guidance generated greater hits than external feature, $t(1033) = 3.00$, $p < .05$, Cohen's $d = .21$, and no guidance conditions, $t(1095) = 2.39$, $p < .05$, Cohen's $d = .16$, while the external feature *vs*. no guidance generated similar performance, $t(1086) < 1$.

There was a prevalence x ability interaction, $F(4, 1583) = 15.13$, $p < .05$, $\eta^2 = .037$. One-way ANOVAs analysed this interaction for each participant group. Results reflected the effects reported above. Hit rates for SRs did not significantly differ across the three prevalence conditions, $F(2, 1583) < 1$. However, hit rates for Experiment 1's typical-ability

controls, $F(2, 1583) = 4.25$, $p < .05$, $\eta^2 = .005$, and with larger effect sizes, Experiment 2's top-end-of-typical range participants significantly differed by condition, $F(2, 1583) = 34.53$, $p < .05$, $\eta^2 = .042$. Controls generated the highest hit rates in the 90% mismatch condition compared to 10% mismatch condition, $t(282.31) = 2.58$, $p < .05$, Cohen's $d = .31$, while other two comparisons did not significantly differ (both *p's* > .05). In direct contrast, Experiment 2 participants generated lowest hits in the 90% mismatch condition compared to both the 50% mismatch, $t(449.21) = 6.16$, $p < .05$, Cohen's $d = .50$, and 10% mismatch conditions, $t(463.06) = 6.60$, $p < .05$, Cohen's $d = .54$, while the latter two conditions did not differ ($p > .05$).

The interaction was also analysed across each mismatch condition. One-way ANOVAs for each were significant, $F_{10\% \text{ mismatch}}(2, 1583) = 34.99$, $p < .05$, $\eta^2 = .042$; $F_{50\% \text{ mismatch}}(2, 1583) = 34.44$, $p < .05$, $\eta^2 = .042$; $F_{90\% \text{ mismatch}}(2, 1583) = 32.00$, $p < .05$, $\eta^2 = .039$. Follow-up t-tests showed SRs outperformed Experiment 2 participants in all conditions (all *t's* ≥ 2.82, $p < .05$). Experiment 2 participants significantly outperformed controls in the 10% and 50% mismatched conditions (both *t's* ≥ 7.66, $p < .05$), but not the 90% mismatch condition, $t(365.05) = 1.35$, $p > .05$.

Other effects and interactions were non-significant (*F*'s < 1.54, $p > .05$)

CRs: This ANOVA revealed a main effect of ability, $F(2, 1583) = 116.51$, $p < .05$, $\eta^2 = .128$. SRs outperformed controls, $t(692.01) = 17.16$, $p < .05$, Cohen's $d = 1.12$, and Experiment 2 participants, $t(963.34) = 10.53$, $p < .05$, Cohen's $d = .57$, while the latter outperformed controls, $t(790.08) = 9.37$, $p < .05$, Cohen's $d = .54$.

There was a main effect of guidance, $F(2, 1583) = 8.54$, $p < .05$, $\eta^2 = .011$. Participants on the no guidance condition made fewer CRs than in the external, $t(1084.12) = 4.09$, $p < .05$, Cohen's $d = .26$, and internal feature guidance conditions, $t(1095) = 2.90$, $p <$

.05, Cohen's d = .18, while the latter two conditions generated similar CRs, $t(1033) = 1.08$, $p > .05$, Cohen's d =.07.

There was a main effect of prevalence, $F(2, 1583) = 5.98$, $p < .05$, $\eta^2 = .007$, though paired comparisons revealed no significant differences ($p > .05$). The non-significant trends were similar to those in Experiment 1, in that 10% mismatch condition generated slightly higher CRs than other two conditions in which CR rates were similar.

There was a prevalence x ability interaction, $F(4, 1583) = 12.78$, $p < .05$, $\eta^2 = .031$. This was analysed separately for each participant group.

CR rates for SRs did not significantly differ across the three prevalence conditions, $F(2, 1583) < 1$. However, CRs for Experiment 1's typical-ability controls, $F(2, 1583) = 20.64$, $p < .05$, $\eta^2 = .025$, and with smaller effect sizes, Experiment 2's top-end-of-typical range participants significantly differed by condition, $F(2, 1583) = 6.08$, $p < .05$, $\eta^2 = .008$. Experiment 1's controls showed lowest CRs on the 90% mismatch compared to 50% mismatch condition, $t(316) = 2.89$, $p < .05$, Cohen's d =.33, and the 10% mismatch condition, $t(230.92) = 4.90$, $p < .05$, Cohen's d = .58, while on this comparison the 10% mismatch condition also generated significantly higher CRs than the 50% mismatch condition, $t(227.09) = 2.75$, $p < .05$, Cohen's d =.32. Experiment 2 participants, on the other hand, showed higher CRs on the 90% mismatch compared to 50% mismatch condition, $t(567) = 3.37$, $p < .05$, Cohen's d = .27, and to 10% mismatch condition, $t(414.15) = 2.81$, $p < .05$, Cohen's d = .25, while the latter two conditions generated similar CR rates ($p > .05$).

Three significant one-way ANOVAs also compared the ability groups in each mismatch condition, $F_{10\% \text{ mismatch}}(2, 1583) = 24.58$, $p < .05$, $\eta^2 = .030$; $F_{50\% \text{ mismatch}}(2, 1583) = 39.55$, $p < .05$, $\eta^2 = .048$; $F_{90\% \text{ mismatch}}(2, 1583) = 83.75$, $p < .05$, $\eta^2 = .096$. Follow-up t-tests showed that SRs outperformed Experiment 2 participants in all three mismatched conditions (all t's ≥ 4.10, $p < .05$). Experiment 2 participants outperformed controls in the 50% and 90%

mismatch conditions (both t's ≥ 6.12, $p < .05$), but not in the 10% mismatch condition, $t(404) < 1$.

Other effects and interactions were non-significant (all F's < 2.08, $p > .05$).

Sensitivity (d'): This ANOVA revealed a main effect of ability, $F(2, 1583) = 272.15$, $p < .05$, $\eta^2 = .256$. SRs outperformed Experiment 2's participants, $t(1156) = 14.80$, $p < .05$, Cohen's $d = .96$, who in turn outperformed controls, $t(1291) = 12.88$, $p < .05$, Cohen's $d = .75$.

There was a main effect of guidance, $F(2, 1583) = 13.93$, $p < .05$, $\eta^2 = .017$. No guidance generated lower performance than external, $t(1086) = 2.40$, $p < .05$, Cohen's $d = .15$, and internal feature guidance, $t(1095) = 4.60$, $p < .05$, Cohen's $d = .28$, while the latter two did not significantly differ, $t(1033) = 2.10$, $p > .05$, Cohen's $d = .13$.

There was a main effect of prevalence, $F(2, 1583) = 19.82$, $p < .05$, $\eta^2 = .032$. Performance was higher in the 10% mismatch condition compared to the 50% mismatch condition, $t(999.42) = 6.89$, $p < .05$, Cohen's $d = .42$, and the 90% mismatch condition, $t(1069) = 4.73$, $p < .05$, Cohen's $d = .29$, while the latter two generated similar sensitivity levels, $t(1047.60) = 1.62$, $p > .05$, Cohen's $d = .10$

There was a prevalence x ability interaction, $F(4, 1583) = 2.53$, $p < .05$, $\eta^2 = .006$. Three ANOVAs examining ability group sensitivity, $F_{\text{SRs}}(2, 1583) = 8.44$, $p < .05$, $\eta^2 = .011$, $F_{\text{controls}}(2, 1583) = 12.38$, $p < .05$, $\eta^2 = .015$, and, $F_{\text{Experiment 2's participants}}(2, 1583) = 11.50$, $p < .05$, $\eta^2 = .014$, across the prevalence conditions were significant. Paired comparisons were conducted.

Sensitivity of SRs in the 10% and 90% mismatch conditions did not differ, $t(205) < 1$, but both were higher than in the 50% mismatch condition (both t's ≥ 3.50, $p < .05$).

In contrast, sensitivity of Experiment 2 participants and controls was highest in the 10% mismatch condition than in the 50% mismatch (both t's ≥ 4.19, $p < .05$), and 90% mismatch conditions (both t's ≥ 3.50, $p < .05$), while the latter two conditions generated similar performance (both t's ≤ 1.53, $p < .05$).

Three one-way ANOVAs analysed ability for each mismatch condition were significant (all F's ≥ 72.07, $p < .05$, $\eta^2 \leq .083$). In all mismatch conditions, SRs significantly outperformed Experiment 2 participants (all t's ≥ 7.69, $p < .05$), who significantly outperformed controls (all t's ≥ 5.86, $p < .05$).

Criterion (C): This ANOVA revealed no main effect of ability, $F(2, 1583) = 1.55$, $p > .05$, $\eta^2 = .002$, but there was a main effect of guidance, $F(2, 1583) = 4.12$, $p < .05$, $\eta^2 = .005$. External feature guidance generated a more conservative response bias (i.e. more likely to respond "different") than internal feature, $t(1026.58) = 2.68$, $p < .05$, Cohen's d =.16, and no guidance, $t(1086) = 2.94$, $p < .05$, Cohen's d =.17, while the latter two generated similar response bias, $t(1095) < 1$.

There was a main effect of prevalence, $F(2, 1583) = 19.06$, $p < .05$, $\eta^2 = .024$. The 10% mismatch condition generated a more conservative response bias than the 50% mismatch, $t(927.87) = 4.44$, $p < .05$, Cohen's d =.27, and the 90% mismatch conditions, $t(1068.25) = 3.54$, $p < .05$, Cohen's d =.21, while the latter two showed similar response bias $t(954.87) < 1$.

There was a prevalence x ability interaction, $F(4, 1583) = 16.55$, $p < .05$, $\eta^2 = .040$.

The SRs, controls and Experiment 2 participants showed different response bias patterns in the 10% mismatch, $F(2, 1583) = 6.76$, $p < .05$, $\eta^2 = .008$, and the 90% mismatch conditions, $F(2, 1583) = 27.32$, $p < .05$, $\eta^2 = .033$, but not in the 50% mismatch condition, $F(2, 1583) = 1.46$, $p > .05$, $\eta^2 = .002$.

In the 10% mismatch condition, controls showed a more conservative response bias than Experiment 2 participants, $t(404) = 2.99$, $p < .05$, Cohen's d =.32, while other comparisons were non-significant ($p > .05$). In the 90% mismatch condition, Controls, $t(456) = 6.26$, $p < .05$, Cohen's d = .61, and SRs, $t(392) = 3.45$, $p < .05$, Cohen's d =.41, showed more liberal (i.e. more likely to respond "same") response bias than Experiment 2 participants, while SRs vs. controls response bias was similar, $t(260) = 1.76$, $p > .05$, Cohen's d = .21

Discussion

In Experiment 2, participants who had achieved scores in the top-end of the typical range on the CFMT+ and GFMT were informed in advance of the approximate prevalence of matched-to-mismatched trials and also completed external, or internal facial feature focus scales, or no scales before making each face matching decision. As expected, and consistent with previous research on low prevalence effects (Weatherford et al., 2020), Experiment 2's participants displayed opposite criterion shift effects to Experiment 1's who were not provided with prevalence information. When informed of mismatched trial infrequency (10% mismatched), participants displayed the typical response bias to respond 'same'. Those provided information that matched trials would be infrequent (90% mismatched), displayed a bias to respond 'different' rather than 'same'. Hit rates were, therefore, higher in the 10%, than the 90% mismatched condition; while, in contrast, CR rates were higher in the 90% than 10% mismatched conditions. These results suggest that participants are less likely to correctly identify low prevalence items, but only if aware of prevalence in advance.

Experiment 2 also showed that external feature focus guidance was more helpful than no guidance in identifying two different identities as mismatched. However, when

Experiment 1's data were combined with Experiment 2 (see supplementary materials), probably a consequence of increased statistical power, internal facial feature focus guidance generated the greatest hits, while CRs and sensitivity were greater for both internal and external feature guidance conditions, compared to no guidance. Interestingly, external feature focus guidance also induced a more conservative criterion shift (i.e., more likely to respond "different") compared to other conditions.

General Discussion

The two experiments described in this paper examined the impact of face recognition ability, facial feature focus guidance, and match-mismatch item prevalence on simultaneous face matching performance. With strong effect sizes, and consistent with the results of previous research (e.g., Russell et al., 2009), Experiment 1's SRs significantly outperformed Experiment 2's top-end of typical participants, who also outperformed typical-range participants. Furthermore, unlike typical-range-controls in Experiment 1, whose CR rates were influenced by match-mismatch trial prevalence, hits and CRs of SRs were not significantly impacted by low prevalence effects. SR's advantage over both sets of controls (Experiments 1 and 2) was found in all prevalence and guidance conditions, albeit not all comparisons within interaction analyses were significant. Therefore, unlike experienced identity verification professionals (Weatherford et al., 2021), individuals with pre-existing superior face recognition ability may be less affected by low prevalence effects. That being said, approximately 10% of SRs achieved maximum scores, therefore it is not possible to rule out ceiling effects as obscuring low prevalence effects in SRs.

The use of feature focus guidance scales, however, to encourage participants to assess the similarity of internal or external features in each pair of images had a far weaker,

albeit positive, effect. In Experiment 1, hits, CRs, and sensitivity were significantly higher in the internal focus guidance condition, than in the no guidance condition. In contrast, external features guidance only improved CRs in comparison to no guidance. As such, the benefits of external features guidance (ears, face shape and jawline) appeared mainly limited to differentiating when two faces were not of the same person. However, similar effects of guidance were not observed in Experiment 2 when participants were informed as to the matched-mismatched trial prevalence, while no interactions were found between the feature guidance conditions and any other condition in either Experiment 1 or 2. As such, the internal feature focus guidance scales provided a significant additional impact on overall accuracy, above that driven by face recognition ability alone, but only when participants were not informed of the relative prevalence of the different trial types.

Experiment 1 revealed an unexpected opposite criterion shift effect to that found in previous low prevalence research, as when participants had no prior knowledge of matched-mismatch prevalence, a decrease in the proportions of mismatched-to-matched items was associated with an unexpected criterion shift so that controls, but not SRs, were more likely to respond ‘different’ or mismatch. Opposite effects, but consistent with most previous research (e.g., Weatherford et al., 2020), were found in Experiment 2 when participants were made aware of prevalence. These findings demonstrate that participants are less likely to correctly identify low prevalence items, but only when aware of likely prevalence in advance.

One explanation for the unexpected effects in Experiment 1, is that all participants had previously taken the GFMT in which 50% of trials are mismatched. Because no trial-type prevalence information was provided, expectations that mismatch-match trial prevalence in Experiment 1 would also be balanced, might have driven decision-making. Indeed, the influence of balanced trial-type expectancy effects would predict Experiment 1’s pattern of results. This would, however, not predict Experiment 2’s results, and it does not appear

credible that GFMT completion generated such strong expectancy effects in Experiment 1 when no prevalence information was provided, which would be entirely negated by the prevalence information in Experiment 2.

An alternative explanation is based on previous face matching results. Weatherford et al. (2020), argue that criterion shifts may be due to participants searching mainly for salient cues in two images that support high-prevalence item expectations, and either ending searches early when their threshold is met, or not attending to, or placing very low weight on any cues denoting low prevalence items. In contrast, the current results suggest that when no information as to prevalence is provided, participants search more equally for salient within-person and between-person cues, and thresholds for both match and mismatch decisions are more equitable.

In addition, as indicated by the significant internal feature guidance scale effects in Experiment 1, but not Experiment 2, it is possible that more weight is placed on scale use which encourages participants to actively search for match and mismatch cues in the image pairs when they have no preconceptions as to likely prevalence. When aware of prevalence, as in Experiment 2, interventions such as scales offer less advantage, as participants search primarily for the cues consistent with their high-prevalence condition expectations. The positive results of Experiment 1, nevertheless, suggest that the use of facial feature guidance scales, particularly those focusing on the internal features of faces, might improve workforce face matching accuracy. Effects were consistent across prevalence conditions and improved performances regardless of participants' face recognition ability. Effects, however, were far weaker than between-face recognition ability groups, and as noted, were not found in Experiment 2. It is perhaps not surprising therefore that previous research investigating similar feature-based interventions on face matching has revealed inconsistent effects at best (e.g., Towler et al., 2017).

Participants in the current study completed three scales only before providing a face matching decision. This was far fewer than the 11 scales completed by those in Towler et al.'s (2017) study, and, it is clear further research is required to identify the key scales, and the number of scales that could improve performances in the workplace. In addition, the images used were chosen from a small database and were primarily selected only to equalise task difficulty across matched and mismatched trials in all prevalence conditions. The properties of images varied within and between each condition, so that for instance, some conditions may have contained more matched images taken with longer delays between image capture than in other conditions. Importantly, relatively stable (i.e., ear size which may substantially increase with long intervals) and unstable (i.e., eyebrows) features may have been more or less impacted by these intervals, and this may also have varied between prevalence conditions. It is not possible to predict how participants may have applied their own knowledge of feature stability over time when using the scales, albeit *within* each prevalence condition, participants in each prevalence condition viewed exactly the same images. It is also entirely plausible that despite using the scales attentively, participants disregarded their scale-based decisions when making their final matching decision to each pair of faces.

In the current research when using scales, participants could also view the full faces, and as such, peripheral information may not have been disregarded and may have driven their decision-making (see, for instance, García-Zurdo, Frowd, & Manzanero, 2020). Obscuring the external or internal facial features entirely can impact strategies used in face matching (e.g., Bruce et al., 1999). Such a procedure might better direct attention to the features of interest. As such, future research could examine whether the use of guidance scales in combination with obscuring features, not of interest might generate stronger effects. Nevertheless, when participants completed no scales, it took approximately 12 min to

complete the 50-trial test. In contrast, participants in the guidance conditions took approximately 30 minutes. As such, organisations would need to decide whether the benefits from the relatively small but significant improvements in accuracy generated by scale use would be outweighed by the increased staff time costs.

Since the initial reports of their skills by Russell et al. (2009), research has consistently demonstrated SR's superior abilities across a variety of face recognition tasks (Bobak et al, 2016; Davis et al, 2016; Robertson et al, 2016). As a result, there has been developing support for the deployment of SRs to identity-critical roles in security and criminal justice operations (Robertson et al., 2016). Here, we showed that those pre-tested to be at the 'top end of the typical range' also significantly outperformed typical-ability controls, albeit the former were provided with the prevalence information that was withheld from the latter. Nevertheless, SRs are rare in the population, and these results suggest that identity verification roles would still be better performed by those just below SR range than those of typical ability. However, consistent with previous research (e.g., Bate, Frowd et al., 2019), Figures 1 and 2 display clear heterogeneous patterns of performance by SRs. Indeed, many controls individually outperformed many SRs across all prevalence and guidance conditions. Nevertheless, poor performances on any task may be a consequence of many factors outside the control of researchers (distractions, fatigue). High scores on the other hand may be a consequence of a series of lucky guesses. Furthermore, beyond a single random prize draw in Experiment 1, no compensation was provided to participants, and motivations to provide accurate responses would likely be lower in research than in the workplace. On the other hand, Weatherford et al. (2021) found no advantage in face matching performance in professional security or ID verifying roles compared to non-professional controls, therefore the lack of professional motivation cannot account for all individual differences observed in

this study. Other face or general processing qualities must contribute to face matching performance variability.

Some researchers have criticised the use of the CFMT+ and GFMT as selection tools for SRs. The GFMT suffers from ceiling effects and has low discriminatory power (Davis et al., 2016), while participants with known face recognition deficits (i.e., prosopagnosia) have been observed to score relatively highly on a shorter version of the CFMT+ (e.g., Esins, Schultz, Stemper, Kennerknecht, & Bulthoff, 2016). This suggests the use of test-specific strategies to achieve high scores may be possible, and that these strategies may be unrelated to face recognition ability. Researchers have therefore suggested the use of recently developed alternative tests (e.g., Dunn et al., 2020), or a battery of tests (e.g., Noyes et al., 2017); for identifying SRs. Nevertheless, the use of multiple tests would substantially increase participant demands. On the other hand, it is likely that between-ability group effect sizes would have increased in the current research if group membership criteria had been more refined.

There were other limitations to the research that should be acknowledged. The images used in the dataset varied in their qualities including some that were up to 10-years apart and accuracy will have been driven by the effects of ageing. However, none of the images depicted children where ageing effects are at their strongest in changing facial appearance, and this should be a focus of future research. Furthermore, participants only completed 50 trials in the current study, chosen as past research using the same database of participants has found dropout rates started to increase after about 30 minutes (the approximate mean time participants took to complete the trials in the feature scale guidance conditions). Some employees working on identity verification tasks may view many more sets of images, and future researchers would be advised to examine effects using more trials.

As with all online studies, there was also less control over conditions than would be possible in a laboratory, although participants were asked not to start unless they were using laptops or personal computers. Nevertheless, with such strong between-ability-group effects, random allocation to all conditions, and large numbers of randomly selected participants from the volunteer database, equipment quality variability is unlikely to explain the effects found.

Conclusions

The combined results of Experiment 1 and 2 suggest that deployment of SRs to identify verification critical roles may have a positive impact on the identity verification workplace. Indeed, recent research suggests that SRs would be useful in policing and identity verification roles as they are less impacted by face occlusions (e.g., face masks and sunglasses, Noyes, Davis, Petrov, Gray, & Ritchie, 2021) whereas interactive image matching procedures developed to aid face matching enhance SR's performance even further (Smith et al., 2021). Although effect sizes were far smaller, the use of internal facial feature guidance scales might also provide an additional advantage, albeit further research is required to define which scales may be most appropriate to employ in different workplace environments. Importantly, supporting previous research evaluating best workplace practices, the advantage of employing SRs over controls, and using internal guidance scales over no guidance was consistent regardless of low prevalence item conditions.

Acknowledgements

The authors would like to thank research assistant Katie Read for her contribution to data collection.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Statement

The authors received no financial support for the research, authorship, and/or publication of this article.

Open Practice Statement

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/n75ha/

References

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision, 14,* 563-563. DOI: 10.1167/14.10.563

Alenezi HM & Bindemann M (2013). The effect of feedback on face matching accuracy. *Applied Cognitive Psychology, 27,* 735–753. https://doi.org/10.1002/acp.2968

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502–508. https://doi.org/10.1111/opo.12131

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications, 3*(1). DOI: 10.1186/s41235-018-0114-7

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face

recognition. *Cognitive Research: Principles and Implications*, *3*(1). DOI: 10.1186/s41235-018-0116-5

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex*, *108*, 92-111. DOI: 10.1016/j.cortex.2018.07.008

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

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. DOI: 10.1037/a0021893

Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4):552-557. doi: 10.7334/psicothema2016.383. PMID: 29048317.

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

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS one*, *11*(2), e0148148.

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

Bobak, A. K., 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*(2), 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*(4), 339. ISSN 1076-898X

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286-291. DOI: 10.3758/BRM.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248. https://www.jstor.org/stable/40063419

Davis, J. P. (2019). The worldwide impact of identifying super-recognisers in police and business. *The Cognitive Psychology Bulletin; Journal of the British Psychological Society: Cognitive Section*, *4*, 17-22. ISSN: 2397-2653. https://shop.bps.org.uk/the-cognitive-psychology-bulletin-issue-4-spring-2019

Davis, J. P., & Robertson, D. J. (2020). Capitalizing on the super-recognition advantage: a powerful, but underutilized, tool for policing and national security agencies. *The Journal of The United States Homeland Defence and Security Information Analysis Center (HDIAC)*, *7*(1), 20-25.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482-505. DOI: 10.1002/acp.1490

Davis, J. P., Bretfelean, D., Belanova, E., & Thompson, T. (2020). Super-recognisers: face recognition performance after variable delay intervals. *Applied Cognitive Psychology, 34*(6), 1350-1368. DOI:10.1002/acp.3712

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. https://doi.org/10.1002/acp.3405

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

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. https://doi.org/10.1111/bjop.12103

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

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th edition). SAGE Publications.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science, 18*(11), 943-947. doi: 10.1111/j.1467-9280.2007.02006.x.

Fussey, P., & Murray, D. (2019). *Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology*. Essex University's Human Rights Centre, retrieved from https://www.hrbdt.ac.uk/download/independent-report-

on-the-london-metropolitan-police-services-trial-of-live-facial-recognition-technology/

Fysh, M.C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology, 109*, 219-231. https://doi.org/10.1111/bjop.12260

García-Zurdo, R., Frowd, C. D., & Manzanero, A. L. (2020). Effects of facial periphery on unfamiliar face recognition. *Current Psychology, 39*, 1767-1773. DOI: 10.1007/s12144-018-9863-1

Gentry, N. W. (2019). *Unfamiliar face matching: Decision-making and improvement* (Doctoral dissertation, University of Kent).

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

Jenkins, R., White, D., Van Montfort, X., Burton, M. A. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323. doi: 10.1016/j.cognition.2011.08.001.

Kemp, R. I., Caon, A., Howard, M., & Brooks, K. R. (2016). Improving unfamiliar face matching by masking the external facial features. *Applied Cognitive Psychology*, *30*(4), 622-627. DOI: 10.1002/acp.3239

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. https://doi.org/10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

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

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PloS one*, *13*(3), e0193455.

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

Moore, R. M., & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology*, *27*(6), 754-760. https://doi.org/10.1002/acp.2964

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*, *27*(12), 1135–1137. https://doi.org/10.1038/nbt1209-1135

Noyes, E., Davis, J. P., Petrov, P., Gray, K. L. H., Ritchie, K. (2021). The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science*, *8*, 201169. https://doi.org/10.1098/rsos.201169

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics*, *76*(5), 1335-1349. doi: 10.3758/s13414-014-0630-6.

Papesh, M. H., Heisick, L. L., & Warner, K. A. (2018). The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied, 24*(3), 416. doi: 10.1037/xap0000156.

Ramon, M., Bobak, A. K., & White, D. (2019). Towards a ‘manifesto’ for super-recognizer research. *British Journal of Psychology*, *110*(3), 495-498. https://doi.org/10.1111/bjop.12411

Robertson, D. J. (2018). Face recognition: Security contexts, super-recognizers, and sophisticated fraud. *The Journal of The United States Homeland Defence and Security Information Analysis Center (HDIAC), 5*(1), 6-10.

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

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PloS One*, *11*(2), e0150036.

Robertson, D. J., Fysh, M. C., & Bindemann, M. (2019). Face identity verification: Five challenges facing practitioners. *Keesing Journal of Documents & Identity*, *59,* 3-8. ISSN 1871-272X

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252-257. doi: 10.3758/PBR.16.2.252

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

Smith, H. M. J., Andrews, S., Baguley, T., Colloff, M. F., Davis, J. P., White, D., Rockey, J. C., & Flowe, H. D. (2021). Performance of typical and superior face recognisers on a novel interactive face matching procedure. *British Journal of Psychology* DOI: 10.1111/bjop.12499

Stephens, R.G., Semmler, C., & Sauer, J.D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied, 23*, 336–353. Doi: 10.1037/xap0000130.

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.

Towler, A., Kemp, R. I., & White, D. (2017). Unfamiliar face matching systems in applied settings. *Face processing: systems, disorders and cultural differences. New York: Nova Science Publishing, Inc*.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47-58. doi: 10.1037/xap0000108.

Weatherford, D. R., Erickson, W. B., Thomas, J., Walker, M. E., & Schein, B. (2020). You shall not pass: how facial variability and feedback affect the detection of low-prevalence fake IDs. *Cognitive Research: Principles and Implications*, *5*(1), 1-15. doi: 10.1186/s41235-019-0204-1.

Weatherford, D. R., Robertson, D., & Erickson, W. B. (2021). When experience does not promote expertise: security professionals fail to detect low prevalence fake IDs. *Cognitive Research: Principles and Implications, 6,* 1-27. https://doi.org/10.1186/s41235-021-00288-z

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PloS ONE*, *10*(10), e0139827.

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

White, D., Phillips, J.P., Hahn, C. Hill, M., & O'Toole, A. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings. Biological Sciences / The Royal Society. 282*. DOI:10.1098/rspb.2015.1292.

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

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*(4), 623-638. doi: 10.1037/0096-3445.136.4.623

Zeinstra, C., Veldhuis, R., & Spreeuwers, L. (2016, September). Discriminating power of FISWG characteristic descriptors under different forensic use cases. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)* (pp. 1-7). IEEE.

Supplementary materials

Experiment 1

RTs: Raw and standardised (z-scores) mean RTs are reported in Tables S1A and S1B, while analyses reported below were conducted on standardised RTs only.

A four-way 2 (Matching: Match trials, Mismatch trials; repeated measures) x 2 (Ability: SRs, controls) x 3 (Prevalence: 10%, 50%, 90% mismatch) x 2 (Guidance: external feature, internal feature) mixed ANOVA analysed mean standardized RTs to complete the three Likert feature rating scales and the identity matching decision. The no guidance conditions were excluded from analyses, as participants provided a single same/different decision only. Accuracy was not included in these calculations due to low incorrect response rates. Accuracy was analysed in Experiment 2.

Overall, SRs generally spent more time on matched trials than controls, whereas controls spent less time on matched trials in the condition with fewer match trials (90% *vs.* 50% mismatch prevalence).

There were no significant main effects, ($F's \leq 3.35$, $p > .05$), but there was a matching x ability interaction, $F(1, 483) = 9.65$, $p < .05$, $\eta^2 = .020$. SRs showed longer RTs than controls on the matching trials, $F(1, 483) = 7.45$, $p < .05$, $\eta^2 = .015$, but not mismatching trials, $F(1, 483) < 1$.

There was also a significant matching x ability x prevalence interaction, $F(2, 483) = 6.74$, $p < .05$, $\eta^2 = .027$. One-way ANOVAs examined prevalence effects on matched and mismatched trials in SRs and controls separately.

SRs displayed no main effects of prevalence on either matched, $F(2, 203) = 2.06$, $p > .05$, $\eta^2 = .020$, or mismatched trials, $F(2, 203) < 1$.

Controls showed a main effect of prevalence on matched trials, $F(2, 280) = 4.81$, $p < .05$, $\eta^2 = .033$, but not mismatched trials, $F(2, 280) < 1$. On matched trials, 90% mismatch

condition responses generated shorter RTs than the 50% mismatch condition, $t(175.62) = 3.03$, $p < .05$, Cohen's $d = .43$, while other comparisons were not significant ($p > .05$).

Other interactions were not significant ($F's \leq 1.71$, $p > .05$).

Table S1A.

Mean raw RT (sec) on each condition in SRs (n = 209) and controls (n = 286)

		SRs		Controls	
Mismatch Prevalence	Feature Guidance	Mean	SD	Mean	SD
		Matched trials			
10%	External	32.64	15.60	27.26	12.95
	Internal	27.86	15.58	29.01	14.18
50%	External	33.16	18.32	32.88	17.62
	Internal	28.33	11.89	30.04	16.92
90%	External	38.72	27.04	25.70	11.10
	Internal	35.38	34.51	24.50	13.21
Total		32.33	21.63	28.21	14.69
		Mismatched trials			
10%	External	28.41	22.96	33.09	39.51
	Internal	26.74	12.10	30.15	19.79
50%	External	30.14	17.78	27.79	13.03
	Internal	29.14	19.37	27.61	14.46
90%	External	29.61	15.06	30.09	16.24
	Internal	23.67	8.77	26.14	9.50
Total		27.93	16.50	29.03	20.54

Table S1B.

Mean RT (z scores) on each condition in SRs (n = 209) and controls (n = 286)

		SRs		Controls	
Mismatch Prevalence	Feature Guidance	Mean	SD	Mean	SD
		Matched trials			
10%	External	0.15	0.86	-0.15	0.72
	Internal	-0.12	0.86	-0.05	0.79
50%	External	0.18	1.02	0.16	0.98
	Internal	-0.09	0.66	0.00	0.94
90%	External	0.49	1.50	-0.24	0.62
	Internal	0.30	1.91	-0.30	0.73
Total		0.13	1.20	-0.10	0.81
		Mismatched trials			
10%	External	-0.01	1.21	0.24	2.09
	Internal	-0.10	0.64	0.08	1.05
50%	External	0.08	0.94	-0.04	0.69
	Internal	0.03	1.02	-0.05	0.76
90%	External	0.06	0.80	0.08	0.86
	Internal	-0.26	0.46	-0.13	0.50
Total		-0.03	0.87	0.02	1.09

Note: analyses were conducted on standardized (z-scores) data.

Experiment 2

RTs: Raw and standardised (z-scores) mean RTs are reported in Tables S2A and S2B, while analyses reported below were conducted on standardised RTs only. The no guidance condition was excluded from analyses. A four-way 2 (Matching: Match trials, Mismatch trials; repeated measures) x 2 (Accuracy: Correct, incorrect; repeated measures) x 3 (Prevalence: 10%, 50%, 90% mismatch) x 2 (Guidance: external feature, internal feature) mixed ANOVA analysed mean standardized RTs to complete the three Likert feature rating scales and the identity matching decision. The no guidance conditions were excluded from analyses, as participants provided a single same/different decision only.

There were no significant main effects and no interactions ($Fs \leq 2.46$, $p > .05$) other than a matching x accuracy x prevalence interaction, $F(2, 298) = 7.73$, $p < .05$, $\eta^2 = .049$.

In the 10% mismatch condition, there were no main effects of accuracy or matching (*F's* < 1, $p > .05$), but there was a significant interaction, $F(1, 60) = 11.70$, $p < .05$, $\eta^2 = .163$. Mismatched trials generated roughly similar RTs for correct and incorrect responses, $t(72) = 2.22$, $p > .05$, $d = .26$, while matched trials generated faster RTs for correct than incorrect responses, $t(158) = 3.27$, $p < .05$, $d = .25$.

There were no significant effects or significant interaction in the 50% mismatch condition (*F's* ≤ 3.29, $p > .05$).

In the 90% mismatch condition, as with the 10% mismatch condition, there was only a significant accuracy x matching interaction, $F(1, 107) = 3.99$, $p < .05$, $\eta^2 = .036$, with other effects non-significant (*F's* < 1, $p > .05$). Unlike in the 10% mismatch condition, mismatched trials generated roughly similar RTs for correct and incorrect responses, $t(150) = 2.28$, $p > .05$, $d = .19$, while matching trials generated faster RTs for incorrect than correct responses, $t(138) = 2.44$, $p < .05$, $d = .21$.

Overall, when match trials were more frequent (10% mismatch prevalence) correct matches were made faster than incorrect matches (false alarms). When match trials were rare (90% mismatch prevalence) correct matches took longer than incorrect matches.

Table S2A.

Mean raw RTs (sec) to correct and incorrect responses in each condition

Mismatch Prevalence	Feature Guidance	Correct Mean	Correct SD	Incorrect Mean	Incorrect SD
			Matched trials		
10%	External	27.06	9.29	41.96	19.71
	Internal	23.40	12.11	42.68	32.44
50%	External	31.01	20.67	43.75	27.15
	Internal	25.59	10.56	39.65	24.50
90%	External	33.82	21.59	39.35	47.81
	Internal	28.86	15.47	39.50	26.92
Total		28.77	16.62	41.02	31.06
			Mismatched trials		
10%	External	29.31	12.14	31.11	17.20
	Internal	27.97	14.50	35.87	43.73
50%	External	27.96	12.77	35.68	23.52
	Internal	27.10	11.72	34.04	17.53
90%	External	28.76	12.68	50.65	46.03
	Internal	27.04	11.13	39.77	24.65
Total		27.88	12.30	38.28	30.38

Table S2B.

Mean RTs (Z scores) to correct and incorrect responses in each condition

Mismatch Prevalence	Feature Guidance	Correct Mean	Correct SD	Incorrect Mean	Incorrect SD
			Matched trials		
10%	External	-0.13	0.57	0.04	0.67
	Internal	-0.36	0.75	0.06	1.10
50%	External	0.11	1.27	0.10	0.92
	Internal	-0.22	0.65	-0.04	0.83
90%	External	0.28	1.33	-0.05	1.62
	Internal	-0.02	0.95	-0.05	0.91
Total		-0.03	1.02	-0.05	1.31
			Mismatched trials		
10%	External	0.09	0.85	-0.27	0.57
	Internal	0.00	1.02	-0.11	1.46
50%	External	0.00	0.90	-0.12	0.78
	Internal	-0.06	0.82	-0.17	0.58
90%	External	0.06	0.89	0.38	1.53
	Internal	-0.06	0.78	0.02	0.82
Total		-0.01	0.86	-0.03	1.01

Note: analyses were performed on standardized (z-scores) data