Published in final edited form as:

Atten Percept Psychophys. 2014 July ; 76(5): 1335–1349. doi:10.3758/s13414-014-0630-6.

Infrequent identity mismatches are frequently undetected

Megan H. Papesh and

Louisiana State University, Baton Rouge, LA 70803, USA

Stephen D. Goldinger

Arizona State University, Tempe, AZ 85287, USA

Megan H. Papesh: mpapesh@lsu.edu

Abstract

The ability to quickly and accurately match faces to photographs bears critically on many domains, from controlling purchase of age-restricted goods to law enforcement and airport security. Despite its pervasiveness and importance, research has shown that face matching is surprisingly error prone. The majority of face-matching research is conducted under idealized conditions (e.g., using photographs of individuals taken on the same day) and with equal proportions of match and mismatch trials, a rate that is likely not observed in everyday face matching. In four experiments, we presented observers with photographs of faces taken an average of 1.5 years apart and tested whether face-matching performance is affected by the prevalence of identity mismatches, comparing conditions of low (10 %) and high (50 %) mismatch prevalence. Like the low-prevalence effect in visual search, we observed inflated miss rates under low-prevalence conditions. This effect persisted when participants were allowed to correct their initial responses (Experiment 2), when they had to verify every decision with a certainty judgment (Experiment 3) and when they were permitted “second looks” at face pairs (Experiment 4). These results suggest that, under realistic viewing conditions, the low-prevalence effect in face matching is a large, persistent source of errors.

Keywords

Unfamiliar face matching; Low-prevalence effect; Signal detection

In August 2012, an amusing story was reported regarding a tour bus in Iceland (Recinto, 2012). After the bus made one of its scheduled stops, a woman (described as “Asian, about 160 cm tall, speaks English well”) had apparently gone missing, having stepped off the bus and never returned. This initiated an intense search that lasted an entire weekend, with approximately 50 people (both tour passengers and police) searching the area, and a helicopter standing by. As it turned out, one of the searchers was the “missing” woman herself! During the tour stop, she had changed her clothes and rejoined the tour, but nobody recognized her. After 2 days of searching, she ultimately realized that the description of the missing woman may have referred to her, and she reported herself “found.” This anecdote

demonstrates the unique challenge of unfamiliar face matching, which is difficult when two photos are compared (e.g., Megreya & Burton, 2008) and may reduce to guesswork when the given information is a crude sketch or vague description. In the present study, we examined the former case—matching two different photos of the same person—with special focus on the ability of people to detect mismatching faces (e.g., fake IDs) when they rarely occur.

The ability to match faces to photographs bears critically on everyday legal and security concerns. The to-be-matched individual is typically unfamiliar to the person verifying their documents, as when customs agents check passports for hundreds of travelers. In such circumstances, however, face-to-photo matching has proven quite fallible. Under optimal viewing conditions in the laboratory, face matching is surprisingly error prone, with error rates between 10 % and 20 % (Bindemann, Avestisyan, & Blackwell, 2010; Megreya, Bindemann, & Havard, 2011). More naturalistic studies have documented even poorer performance (Kemp, Towell, & Pike, 1997), suggesting that error rates in applied situations may be alarmingly high.

When individuals encounter familiar faces (i.e., individuals with whom we have prior perceptual experience; Hancock, Bruce, & Burton, 2000), they are able to quickly and accurately identify them, despite large perceptual and contextual changes, such as changes in lighting, viewpoints, or facial hair (Burton, Wilson, Cowan, & Bruce, 1999). These perceptual compensations do not typically extend to unfamiliar faces. Instead, unfamiliar face matching is a complex signal detection process: Observers must extract diagnostic identity information from noisy and potentially variable signals (i.e., the live face and the photo ID), while adopting a criterion for whether that information signals a match or a mismatch. The “search” process involved in face matching is largely underspecified. Unlike many perceptual search tasks, which are typically guided by categorical targets (e.g., “find a weapon” or “detect a tumor”), face matching uses imprecise target templates; observers cannot anticipate a predefined feature or characteristic that will signal identity matches or mismatches (Burton, Jenkins, Hancock, & White, 2005; Megreya & Burton, 2006). Indeed, to perform the task optimally, observers must tolerate substantial differences across images, still classifying them as the same person. At the same time, those same feature changes also signal different people, which makes face matching a “textbook” signal detection challenge, potentially subject to large effects of criterion changes.

Prior research on the difficulty of identifying unfamiliar faces has focused on the role of memory in the identification process, particularly in the context of eyewitness memory (e.g., Lindsay & Pozzulo, 1999; Malpass & Devine, 1981; Wells & Olson, 2003). In such research, there is an implicit assumption that misidentification errors arise because observers must match perceptual input (e.g., a person in a lineup) to events in long-term memory that may have degraded over time. Recent research, however, has suggested that unfamiliar face identification is surprisingly error prone, even when memory demands are minimized (or eliminated), as in matching tasks. Laboratory matching tasks are essentially “bare bones” versions of the task faced by security agents and sales clerks everyday: Participants are shown two photographs (or a live actor and a still photograph) and must decide whether they are an identity match or a mismatch. Using this paradigm, Hill and Bruce (1996) discovered

the surprising fallibility of unfamiliar face matching: With unlimited time and no memory demands, participants' ability to match two pictures of the same face dropped substantially when the faces were depicted from different viewpoints or in different lighting conditions. Because the flattering fluorescent lighting of the Motor Vehicles Department is rarely replicated when ID verification takes place at passport control, or during the purchase of age-restricted goods, research into the process of unfamiliar face matching is vitally important.

Whereas Hill and Bruce (1996) found poor face-matching performance under mildly compromised perceptual conditions (changes in viewpoint and lighting), Bruce et al. (1999) reported that unfamiliar face matching remains difficult even under seemingly ideal conditions. In their study, participants attempted to match a still photograph of a young man (the target) to another photo of the same man in a 10-face lineup. Their first experiment included a condition in which target and lineup photos were all taken on the same day, from the same viewpoint, and with good lighting. The cameras used to take the photographs, however, differed (one was taken from a high-quality video, and the other was taken with a studio film camera). Even under these nearly ideal conditions, with no time constraints and an emphasis on accuracy, participants made approximately 30 % errors in both target-present and target-absent displays. (In other conditions, Bruce et al., 1999, replicated the detrimental effects of changes in viewpoints or expressions.) Error rates dropped slightly (to ≈ 20 %) when all trials contained target-present lineups. These poor performance levels did not change when the photographs were in color or grayscale, or when participants watched the original video of the target before viewing the lineup. This result has been replicated extensively (e.g., Bruce, Henderson, Newman, & Burton, 2001) and with different stimulus sets (e.g., Megreya & Burton, 2007).

Whereas Bruce et al. (1999) examined matching ability with face lineups, other researchers have used just pairs of faces (Clutterbuck & Johnston, 2002; Megreya & Burton, 2006, 2007), which is similar to ID verification. Investigating individual differences in face-matching ability, Megreya and Burton (2006) observed that performance using face pairs (82 % hits) was no different than performance with 10-face lineups (≈ 80 % hits). In every case, viewers made between 10 % and 25 % errors. Such high error rates would be unacceptable for a clerk verifying ages for alcohol purchases; they would be devastating for someone charged with ensuring airline security. One key difference, however, between laboratory face-matching studies and real-life security situations is that laboratory studies often use static photo-to-photo matching, whereas security situations involve matching live individuals to static photographs.

To investigate ID verification as it more naturally occurs and whether the use of photo-to-photo matching in the laboratory is justified, some researchers have examined face-matching performance using live targets and static photographs. Davis and Valentine (2009) investigated mock jurors' abilities to match a live suspect to a CCTV image of that person (or someone else) committing a crime. When the CCTV footage was recorded 3 weeks prior to the matching task, participants incorrectly responded "mismatch" 22 % of the time, and they incorrectly identified an innocent suspect 18 % of the time. These error rates increased when the footage was taken a year prior and when the suspect had changed surface-level

features (e.g., glasses or facial hair) during the intervening time. Other researchers have observed similar results under more naturalistic conditions. For a while, U.K. authorities considered adding cardholders' photographs to the fronts of their credit cards, to combat credit card theft (indeed, several credit card issuers adopted this policy on their own). Kemp et al. (1997) examined the utility of this approach, sending shoppers through a supermarket to purchase goods using such credit cards. Some of the shoppers paid with cards containing their own picture, taken no more than 6 weeks prior; others paid with another individual's credit card (matched for race and sex). Because the cards were kept in opaque envelopes, shoppers were blind to whether they were shopping with real or "stolen" credit cards. Cashiers, however, were aware of the ongoing study and were explicitly warned to check the photographs. Despite this, the cashiers accepted fraudulent cards for more than 50 % of the transactions. In both idealized laboratory conditions and naturalistic studies, matching live individuals to static photographs yields surprisingly poor performance ($\approx 15\%$ errors), which is no different than matching two static photographs (Megreya & Burton, 2008).

To investigate unfamiliar face matching, most studies utilize equal proportions of match and mismatch trials (although several lineup-based studies have included conditions in which all lineups are target present). Real-life photo ID matching rarely provides such odds, particularly at security checkpoints or passport control, where the probability of being presented with a fake ID is undoubtedly quite low. There is strong evidence to suggest that the infrequency with which individuals encounter fraudulent IDs will have a large impact on their ability to detect such IDs. This is known as the *low-prevalence effect* (LPE) and is commonly observed in perceptually challenging visual search tasks wherein targets occur only rarely. In LPE studies, researchers commonly observe that, as target prevalence decreases, so too does the ability to detect targets. Wolfe, Horowitz, and Kenner (2005) manipulated the prevalence of targets (tools) among nontool distractors embedded in noisy backgrounds, using target-present rates of 50 %, 10 %, and 1 %. Whereas observers missed only 7 % of targets in the high-prevalence (henceforth HP) condition, they missed 30 % in the low-prevalence (LP) condition. Miss rates of 30 % are harmless in the laboratory, but in real contexts (e.g., airport baggage screening, radiology, military image analysis), observers cannot afford to miss 30 % of potential threats.

Wolfe et al. (2005) suggested that the LPE occurs because of context-based criterion shifts: Observers become conservatively biased under LP conditions, terminating searches more quickly and missing more targets. Fleck and Mitroff (2007), however, proposed that observers may not truly "miss" the targets but may, instead, respond too quickly, using a prepotent motor response that gets initiated prior to conscious awareness of the target. To examine this hypothesis, they gave participants the option to change their initial search decisions (from "target absent" to "target present" and vice versa). By analogy, if a baggage screener suspects that a quickly glimpsed object in a carry-on bag may have been a knife, she will rerun the bag through the x-ray machine or will pull it aside for further inspection. In such applied contexts, decisions are not speeded, dichotomous choices with no recourse under conditions of uncertainty; observers are allowed to reevaluate decisions as they deem necessary. By comparing decisions before and after participants corrected their responses, Fleck and Mitroff observed a drop in the LPE from 27 % to 10 %. In fact, although

participants could use the “correction key” on any trial, they used it to correct target misses over 90 % of the time, supporting the argument that the LPE is, at least partially, an error of motor execution in laboratory search tasks. This does not seem to be the entire explanation, however, as the LPE has proven stubborn to eliminate in other contexts (e.g., when observers are required to remake any decision that was originally made too quickly; Wolfe et al., 2007).

Although the LPE is robust in visual search, the apparent cause of the effect differs, on the basis of task demands. Rich et al. (2008) observed a prevalence effect in perceptually simple feature search (seeking a rotated line among distractors) and also in a more challenging, spatial configuration search (spotting a T among offset Ls). The LPE in feature search could be eliminated by enforcing a minimum response delay, suggesting that the error was driven largely by a motor component. Conversely, the LPE in configuration search was unaffected by such response manipulations. Rather, these searches were often characterized by observers terminating search too early, which was accompanied by reduced eye movements around the display. Additionally, perceptually challenging visual search might introduce ambiguous targets that are difficult to discriminate from distractors; in this case, observers might adopt a conservative response bias, making more miss errors (Rich et al., 2008). Given that low target prevalence is a persistent source of errors in many types of visual search, with multiple potential underlying causes, the present study assessed whether it would similarly affect face matching, a task that also requires searching visual displays for evidence that a “target” (i.e., an identity match) is truly present. Because the degree of visual similarity between two photos of the same person can vary greatly across instances, it seems likely that if mismatches occur very rarely, observers will become increasingly tolerant of slight differences in features or will make more cursory decisions.

To date, little research has addressed this hypothesis, but it has practical and theoretical importance. The practical importance of face matching is clear and wide-ranging. On the theoretical side, discovering the LPE in face matching would carry at least two implications. First, we suggest that face matching is a classic signal detection problem: People change in appearance over time, with changes in hairstyles, facial hair, eyewear, emotional expressions, and so forth. Collectively, some pairs of matching photos will present stronger “signals” than others (e.g., see Fig. 1), and most are difficult to characterize as simple feature search. By reducing the proportion of “false IDs,” we may encourage observers to either truncate their search for mismatching features or relax their criterion for classifying matches (cf. Rich et al., 2008).

Second, even if mismatches are rare, few perceptual theories would anticipate observers to look at two clearly mismatching faces—shown side by side—and respond “match.” Should the LPE occur in the present study, it would appear to connect unfamiliar face matching with the broader literature showing powerful effects of expectations and top-down matching in visual search. In a well-known study of *change blindness*, Simons and Levin (1998) showed that, in a live conversation with a stranger, people rarely noticed when their interlocutor was exchanged with another person, following a brief visual occlusion. In such a case, all contextual cues (and a lifetime of experience) suggest that people do not change into other people mid-conversation; these top-down factors result in observers “seeing”

matches that do not exist. Although people who screen IDs for a living must be aware that occasional fakes are encountered, their low prevalence might similarly encourage “false seeing.” Although the face-matching task is described as being memory free (Megreya & Burton, 2008), observers must look at one depicted face, then another, in separate fixations. We suggest that people must encode the first into visual working memory, then compare that memory with the adjacent face. Each redirection of gaze will overwrite the contents of working memory, necessitating comparisons of memory with perception. When context strongly suggests that the faces will be the same person, such top-down matching may induce change blindness, in a way that would not occur with faces of familiar people.

To date, only one study has evaluated the LPE in face matching: Bindemann et al. (2010) examined conditions of low (2 %) or high (50 %) mismatch prevalence. In their study, observers completed two blocks of 50 trials (one LP, one HP) and were informed of the prevalence prior to the first trial. In the LP condition, only one face pair was an identity mismatch, and it always occurred on the final trial. Under these conditions, Bindemann et al. observed no LPE; detecting mismatches was just as likely in both conditions. Across four experiments, using both unlimited and time-limited viewing, they observed no effects of mismatch prevalence on performance, although matching performance was generally low, replicating previous studies. It is difficult, however, to draw parallels between Bindemann et al.’s experiments and other visual search studies (or applied situations), given that participants were informed of the prevalence manipulation and only one mismatch trial was used in the LP condition. Visual search experiments, wherein the LPE is robust, typically involve many trials, such that a subset can be target present, even under LP conditions. Also, because they focused on face matching under ideal conditions, Bindemann et al. used face pairs in which the photographs were taken on the same day, approximately 15 min apart (albeit with different cameras).

In the present study, we examined face matching under highly variable conditions more closely related to ID verification and to standard visual search. We hypothesized that, given the perceptually challenging nature of visual search tasks that typically yield the LPE, the same effect would be observed in a realistic, challenging face-matching task. We made the task challenging by using over 200 trials, and embedding target faces into realistic (false) driver’s licenses. Furthermore, our face stimuli were selected with natural variation, in a manner that would closely simulate ID verification in real-world settings. To accomplish this, we first collected images of consenting volunteers from their online university roster photographs: These were school ID photos, taken under constant conditions, similar to government-issued ID cards. We then used a different camera to take an additional photograph of each student under nonconstant lighting conditions. All students in the stimulus set were from Arizona State University, whereas experimental participants were from Louisiana State University, making it highly unlikely that any participants would be accidentally familiar with any depicted individuals.

In our final stimulus set, the age difference between the ID photos and our new photos ranged from a few months to 7 years, with an average difference of approximately 1.5 years. Finally, we presented these faces in a manner to simulate photo-ID matching, such that one photo approximated the size of a human face and the comparison photo was scaled to the

size of a standard ID photo. By constructing our items in this manner, we conceptually equated our face-matching task with categorically guided visual search, wherein observers do not have precise, 1:1 mappings of mental templates to search targets (e.g., a radiologist knows to look for a tumor, but not “this specific tumor”). Although this procedure contains no ambiguity about the locations of the comparison faces and, thus, seemingly differs from visual search, there is trial-to-trial variation in both the location and quality of diagnostic cues within the faces. Given these modifications, relative to Bindemann et al. (2010), we expected to observe a robust LPE in face matching.

General method

All experiments followed the same basic method; any deviations are noted on a per-experiment basis.

Stimuli

The stimuli consisted of paired photographs of consenting volunteers from Arizona State University. One photo in each pair was a standard student ID photo; the other photo was taken with a 7-megapixel digital camera under variable lighting conditions (but with a consistent, light blue background), up to 7 years after the original student ID photo was taken. As is shown in Table 1, the depicted students represented a naturally varying range of (self-reported) ethnic identities. Given the varying time lapses between photos, it was possible for substantial changes to occur, allowing for natural within-person variability, as depicted in columns A and B of Fig. 1. For each person, there was one matching photo and one hand-selected mismatching foil photo (Fig. 1, column C). Foils were selected to be individuals of the same sex and race, without large differences that would make mismatch detection trivially easy. However, as is shown in the figure, detecting mismatches was certainly feasible. Whenever possible, we selected foils from a pool of 46 student ID photographs for which we did not have recent photographs available; because of this limited pool, some student ID photos were used more than once in an experiment. Large, recent photos were never repeated, since they would be more salient and memorable. Across participants, faces were shown with their matching and mismatching counterparts equally often.

ID photos were embedded into realistic “fake IDs” made in Microsoft Powerpoint®; each face was given a unique (false) name, address, birthdate, ID number, and other details. Stimuli were displayed on a $1,920 \times 1,080$ pixel, 21.5-in. LCD Dell monitor. IDs were sized to 366×234 pixels (5.08×3.25 onscreen inches), which was the approximate size of a driver’s license on the computer screen (see Fig. 2). Recent photos were sized to fit within 575×600 pixels, approximately the size of a live face, subtending a visual angle of 9.61° .

Procedure

After providing informed consent, participants were told that their primary task would be ID verification, similar to what is done at the airport or the grocery store with age-controlled items (although with no requirement to check dates of birth). After two practice trials, participants were informed of the penalties for incorrect responses, which were modestly

scaled to the real-life gravity of ID-matching errors: For false alarms (which we defined as incorrectly responding “mismatch” to matching faces, as in columns A and B in Fig. 1), they were penalized with a 2-s pause before the procedure continued, but for misses (defined as responding “match” to mismatched faces, as in columns B and C in Fig. 1), they were penalized 4 s. This was intended to reflect the more serious nature of a miss, relative to a false alarm, in real life. The slightly higher penalty for misses was also intended to help inoculate participants to the LPE, since they would be motivated to avoid misses if possible.¹ Participants then completed 242 face-matching trials, indicating their decisions by pressing the “f” or “j” keys, with the response mappings counterbalanced across participants. In the LP condition, 24 trials (10 %) were mismatches; in the HP condition, 121 (50 %) were mismatches.² Neither the participants nor research assistants were informed of prevalence group assignments.

Experiment 1

Experiment 1 was conducted to examine the effects of mismatch prevalence on unfamiliar face matching, with no time constraints and only a single “match/mismatch” decision made on every trial.

Method

Participants—Sixty-one Louisiana State University students participated for partial course credit. By random assignment, 30 participants ($M_{\text{age}} = 20.8$, 24 females) were assigned to the HP condition, and 31 ($M_{\text{age}} = 19.9$, 26 females) were assigned to the LP condition. The majority ($n = 48$) of our participants self-identified as White, while 11 self-identified as Black, and 2 as Asian. All participants reported normal or corrected-to-normal vision.

Results and discussion

Across all analyses, alpha was maintained at .05, and multiple comparisons were Bonferroni corrected. We operationally classified responses on the basis of the presumed priority of checking photo IDs in order to catch impostors: Hits were defined as correctly detecting mismatching faces, and misses were defined as failing to detect such mismatches. Correct rejections were defined as correctly detecting matching faces, and false alarms were defined as incorrectly claiming that such faces mismatch.

These terms were used to ground all signal detection analyses reported in this article. As is common in signal detection analyses, perfect performance in any given cell (per individual participant) was adjusted by .001.

Response time (RT) outliers were identified for each individual participant, defined as responses falling more than 2.5 standard deviations from that participant’s mean. Those

¹Although a two-second asymmetry in time penalties is subtle, relative to other potential payoff structures (such as monetary penalties), we did not want to induce strategies wherein participants would merely try to maximize payoffs. Across all experiments, the miss and false alarm rates in the high-prevalence conditions were equivalent, suggesting that the extra penalty for misses did not strongly affect performance.

²Note that what we denote as “low prevalence” would likely be considered “medium” in the visual search literature. Within the context of our experiments, however, we refer to the 10 % mismatching condition as being low prevalence.

outliers were replaced with the cutoff RT; these adjusted values were included in RT analyses (Winer, 1971). One trial from 1 participant was dropped for being too fast; otherwise, all outliers reflected slow responses. Across all four experiments, no more than 4 % of trials were considered outliers, with no difference in proportions across match and mismatch trials. Because instructions to participants emphasized accuracy rather than speed, all adjusted-RT trials were still included in the accuracy analyses.

Face matching—We analyzed misses and false alarms in separate, univariate ANOVAs, examining the factor of prevalence (low, high). Although we report raw values for clarity, all accuracy analyses (in all experiments) were conducted on arcsine square-root transformed data, to ensure normality. As is shown in Fig. 3, miss rates were reliably affected by the frequency with which mismatches occurred, $F(1, 59) = 27.83, p < .001, n^2_p = .32$. Participants in the LP condition made more miss errors ($M = .49, SE = .03$) than those in the HP condition ($M = .24, SE = .03$). Consistent with the LPE in the visual search literature (but with greater magnitude), participants in the LP group failed to detect identity mismatches on nearly half of all mismatch trials; this error rate was almost double that observed in the HP group. Complementing the miss error data, we also observed a prevalence effect in false alarms, since the LP group made fewer false alarms ($M = .11, SE = .01$) than the HP group ($M = .21, SE = .01$), $F(1, 59) = 32.26, p < .001, n^2_p = .35$.

Regarding the observed LPE, one potential concern was that, because the specific mismatching trials (i.e., the photos selected for comparison) were not always the same across the LP and HP conditions, uncontrolled selection artifacts might have skewed the results. For example, with relatively few mismatch trials in the LP condition, a handful of compelling mismatches could have a large effect. To rule out such a possibility, in all experiments, we followed the standard, participant-based analyses with complementary item analyses, focused only on the “mismatch” trials.³ Although mismatching pairs were randomly selected for each session, the vast majority were used in both LP and HP conditions, across participants. We conducted paired-samples *t*-tests, comparing performance under LP and HP conditions, treating mismatching photo pairs as “subjects.” We found no evidence that selection artifacts created the LPE. In Experiment 1, all 242 mismatching pairs were sampled in both the LP and HP conditions, with corresponding miss rates of .52 and .23, $t(241) = 12.9, p < .001$.

Because there were hits and false alarms in each group, we derived the signal detection indices d' (sensitivity) and c (bias), which were analyzed in separate univariate ANOVAs. Like the LPE in visual search (e.g., Wolfe et al., 2007), there was no difference in d' across the LP ($M = 1.61, SE = 0.11$) and HP ($M = 1.66, SE = 0.12$) conditions, $F(1, 59) = 0.09, p = .76$, suggesting that the effect is not the result of changes in observers' sensitivity to the perceptual characteristics of the photographs. Instead, as in the visual search LPE, the prevalence effect was associated with changes in response bias (c), $F(1, 59) = 30.32, p < .05$,

³Other potential remediation approaches for this problem include using a constant subset of mismatching trials across the LP and HP conditions or yoking LP participants to HP participants. Although both solutions would overcome potential selection artifacts, the present item analyses also rule out potential influences of any idiosyncratic face pairs, while allowing us to randomly sample from the full set of stimulus photos.

$n^2_p = .34$; participants in the LP group responded more conservatively ($M = .61$, $SE = .07$) than participants in the HP group ($M = .04$, $SE = .07$).

Response times—Because the LPE is often characterized by faster decisions under LP conditions (e.g., Wolfe et al., 2005), we analyzed RTs in a 2 (trial type: match/mismatch) $\times$ 2 (prevalence: low/high) $\times$ 2 (accuracy) mixed-model repeated measures ANOVA, with prevalence as the between-subjects factor. There was a main effect of accuracy, $F(1, 59) = 24.28$, $p < .001$, $n^2_p = .29$, reflecting shorter RTs during correct trials ($M = 2,061$ ms, $SE = 81$), relative to incorrect trials ($M = 2,390$ ms, $SE = 126$). This, however, was qualified by a reliable three-way interaction, $F(1, 59) = 32.96$, $p < .001$, $n^2_p = .36$. As is shown in the left-hand side of Fig. 4, in the LP condition, people generally made fast “match” decisions (misses and correct rejections), which is consistent with the suggestion from Wolfe et al. (2007; see also Fleck & Mitroff, 2007) that people adopt conservative criteria under LP conditions. Conceptually, this replicates the typical LPE, indicating that observers abandoned their search for diagnostic identity cues in the faces too quickly, thereby failing to detect critical mismatching features. In the HP condition (right-hand side of Fig. 5), people were fast to generate correct decisions (hits and correct rejections), relative to errors. This RT pattern was consistently observed in all four experiments reported in this article. Because our primary focus was on face-matching accuracy, we report all remaining RT means and ANOVAs in the Appendix, omitting them from Results sections, although we revisit the RT results in the General Discussion section.

Experiment 2

Because Experiment 1 revealed a reliable LPE in unfamiliar face matching, an effect that has not previously been observed (cf. Bindemann et al., 2010), we sought to replicate it in Experiment 2, using conditions inspired by Fleck and Mitroff (2007). Like the LPE in visual search, the prevalence effect in Experiment 1 was associated with shorter decision times during LP misses. If the LPE in face matching is the result of response execution errors (i.e., responding too quickly, despite having actually detected identity mismatches), then allowing participants to correct their initial decisions should decrease the miss error rate, as in Fleck and Mitroff. Therefore, the aim of Experiment 2 was to test the hypothesis that the LPE in face matching is a motoric error and is not solely associated with context-sensitive criterion shifts.

Method

We used a correction paradigm similar to that developed by Fleck and Mitroff (2007) for studying the prevalence effect in visual search. For half the participants in Experiment 2 (divided into two *standard* groups; one LP and one HP), the procedure was identical to Experiment 1, such that participants were permitted to make only a single decision per trial. The other half of the participants (divided into two *correction* groups, one LP and one HP), were given a 1,000-ms intertrial interval (ITI) period, during which they were permitted to change their initial response on that trial. The stimuli were not visible during the ITI. For example, if a person’s first response was “mismatch,” pressing the space bar would change that response to “match.” The standard groups had the same ITI, but no keypress was

allowed, nor was any such possibility mentioned. All other procedural aspects were the same as those in Experiment 1.

Participants—Eighty-three Louisiana State University students participated for partial course credit. By random assignment, 39 participants ($M_{\text{age}} = 20.1$, 32 females) were assigned to the standard condition (18 in HP, 21 in LP), and 44 ($M_{\text{age}} = 20.3$, 38 females) were assigned to the correction condition (20 in HP, 24 in LP). Please note that the larger sample, relative to Experiment 1, reflects the addition of two extra participant groups. The majority of participants ($n = 59$) self-identified as White, while 11 self-identified as Black, 11 as Asian, and 2 as Hispanic. All participants reported normal or corrected-to-normal vision.

Results and discussion

As in Experiment 1, error rates in face matching (misses and false alarms) were analyzed in separate 2 (response condition: standard/correction) $\times$ 2 (prevalence) between-subjects ANOVAs. A main effect of prevalence, $F(1, 79) = 64.92, p < .001, n^2_p = .45$, replicated the effect observed in Experiment 1. Participants in the LP groups made significantly more misses ($M = .47, SE = .03$) than did participants in the HP groups ($M = .21, SE = .01$). As is shown in Fig. 5, this effect was not qualified by an interaction with response condition, $F(1, 79) = 0.38, p = .54$. The LPE was equivalent in the standard and correction groups, suggesting that it was not a product of motor errors. As in Experiment 1, we conducted an item analysis, treating mismatching face pairs as subjects: The LPE remained robust, with 45 % errors in the LP condition and 19 % errors in the HP condition, $t(222) = 11.4, p < .001$. We also observed differences in false alarm rates as a function of mismatch prevalence, $F(1, 79) = 74.80, p < .001, n^2_p = .49$. Participants in the HP group made more false alarms ($M = .20, SE = .01$), relative to participants in the LP group ($M = .08, SE = .01$).

To examine whether mismatch prevalence affected sensitivity or bias, the data were transformed into the signal detection indices d' and c , which were analyzed in separate 2 (prevalence) $\times$ 2 (response condition) ANOVAs. As in Experiment 1, there was no difference in d' as a function of prevalence, $F(1, 79) = 3.21, p = .08$, although d' differed as a function of response condition, $F(1, 79) = 7.19, p = .01, n^2_p = .08$. Consistent with the small effect size, participants in the standard group had a small, but reliable, increase in d' ($M = 1.77, SE = 0.09$), relative to participants in the correction group ($M = 1.47, SE = .07$). Replicating Experiment 1, the LPE manifested as a main effect in the bias index c , $F(1, 79) = 116.15, p < .001, n^2_p = .60$; participants in the LP condition responded conservatively ($M = .68, SE = .05$), relative to those in the HP condition ($M = .03, SE = .04$).

Although participants in the correction group were encouraged to use the correction key, only 6 participants took advantage of this (4 from the HP condition and 2 from the LP condition). Although there were not enough data to permit statistical analysis, descriptive statistics are shown in Table 2. Whereas Fleck and Mitroff (2007) observed that participants used the correction key mainly to reverse potential miss errors, we did not conceptually replicate this pattern. Collapsed across the LP and HP groups, participants used the

correction key 8 times on match trials (attempting to correct false alarms) and 9 times on mismatch trials (attempting to correct misses) (Table 3).

Experiment 3

In Experiment 2, we replicated the LPE in unfamiliar face matching and did not observe evidence to suggest that such errors result from response execution problems. Although we gave participants 1,000 ms after every response to indicate whether they would like to change their initial decision, very few took advantage of the opportunity, particularly among participants in the LP group. Because we had so few corrections in Experiment 2, we designed Experiment 3 to force participants to make certainty judgments for all decisions, rather than allow them to passively await the next trial. If the LPE in unfamiliar face matching is the result of motor errors, encouraging participants to reevaluate every decision should eliminate the effect.

Method

The procedures were largely the same as Experiment 2, except that participants were *prompted* on every trial to make a certainty judgment (conversely, in Experiment 2, participants simply pressed the space bar during the ITI to indicate whether they wished to change the preceding response). Following every match/mismatch decision in Experiment 3, the screen displayed the question “Are you sure?” to which participants responded “yes” or “no” by pressing the “z” and “/” keys (with mapping counterbalanced across participants). Participants were informed that, for every “no” response, their decision on the immediately preceding face-matching trial would be reversed. In addition, Experiment 3 did not include a standard group, since it would be entirely redundant with Experiments 1 and 2.

Participants—Fifty-one Louisiana State University students participated for partial course credit. One participant was dropped for failing to follow instructions, leaving 50 for analysis. By random assignment, 22 participants ($M_{\text{age}} = 19.5$, 17 females) were assigned to the HP condition, and 28 ($M_{\text{age}} = 19.75$, 19 females) were assigned to the LP condition. Thirty-four participants self-identified as White, 5 self-identified as Black, 5 as Asian, 5 as Hispanic, and 1 as Native American. All participants reported normal or corrected-to-normal vision.

Results and discussion

Analysis on participants’ final miss rates⁴ as a function of prevalence revealed a reliable LPE, $F(1, 48) = 53.87, p < .001, n^2_p = .53$. Even when prompted to reconsider their decisions after every trial, participants in the LP group made more miss errors ($M = .46, SE = .02$) than did participants in the HP group ($M = .20, SE = .03$). Critically, as is shown in Fig. 6, this effect was also observed using initial miss rates, $F(1, 48) = 38.64, p < .001, n^2_p = .45$; participants in the LP group missed more mismatches ($M = .44, SE = .03$) than did participants in the HP group ($M = .20, SE = .03$). These parallel results further suggest that the LPE in unfamiliar face matching is not a motor error, since all participants were required

⁴Throughout the results of Experiments 3 and 4, when error rates are reported as “final”, they include any response corrections made by participants. “Initial” error rates reflect decisions before any corrections were incorporated.

to make initial decisions and then confirm (or reverse) those decisions. As in prior experiments, we replicated the LPE in an item analysis on mismatching pairs, ensuring that the effect was not a selection artifact: Across observers, the same face pairs produced 49 % misses in the LP condition and 20 % misses in the HP condition, $t(228) = 13.3, p < .001$. We also replicated the differences observed in false alarms as a function of mismatch prevalence. In the final false alarms, participants in the LP group made fewer ($M = .08, SE = .01$) errors, relative to participants in the HP group ($M = .21, SE = .01$), $F(1, 48) = 77.98, p < .001, n^2_p = .62$. The same pattern was observed in the initial false alarm rates, $F(1, 48) = 39.25, p < .001, n^2_p = .45$.

Consistent with the criterion shift explanation for Experiments 1 and 2, we observed that the LPE in Experiment 3 manifested as a change in bias (c), $F(1, 48) = 60.52, p < .05, n^2_p = .56$, but not sensitivity (d'), $F(1, 48) = 0.89, p = .35$. Whereas participants in the HP group did not adopt any bias ($M = -.01, SE = .06$), those in the LP group adopted a strong conservative bias ($M = .65, SE = .06$). This difference was observed whether we analyzed c calculated from initial or final decisions.

To assess whether our response certainty prompt encouraged participants to truly consider (or reconsider) their responses, we counted the frequency of “not sure” responses in both groups. In the HP condition, there were a total of 5,324 trials (22 participants $\times$ 242 trials, with equal matching and mismatching trials (2,662 each). Participants responded “not sure” 78 times on matching trials (a “not sure” rate of 2.93 %) and 140 times on mismatching trials (a rate of 5.26 %). In the LP condition, there was a total of 6,776 trials (28 participants $\times$ 242 trials), with 6,104 matching trials and 672 mismatching trials. Participants responded “not sure” 149 times on matching trials (2.44 %) and 41 times on mismatching trials (6.10 %). Therefore, both the LP and HP groups were more likely to correct responses following mismatching trials, with no clear difference between groups. These results are shown in a different form in Table 4, focusing on the accuracy of those corrected trials, irrespective of relative base rates in the full experiment. Note that, when participants in the LP group responded “not sure” during mismatch trials, they were largely incorrect; their miss rate on “corrected” trials was 79 %, suggesting that they truly failed to detect mismatches and did not just respond with a prepotent motor response encouraged by the infrequency of “mismatch” pairs. This suggests that the LPE in unfamiliar face matching may be more persistent than that observed in visual search.

Experiment 4

Thus far, we have observed a robust LPE in unfamiliar face matching that is not mitigated by allowing participants to correct their initial responses in a manner that is either self-initiated (Experiment 2) or required (Experiment 3). This prevalence effect, therefore, appears more persistent than the effect observed in many visual search tasks: Fleck and Mitroff (2007) observed diminished miss rates when participants were permitted to correct their initial responses (although Van Wert, Horowitz, & Wolfe, 2009, found that some rare targets were missed even when participants had the option to correct their initial responses). In fact, the LPE in standard visual search typically reaches a maximum of ≈ 30 % misses; our miss rate was consistently ≈ 45 % or above, suggesting that unfamiliar face matching

poses a greater perceptual challenge, relative to standard visual search. Because the LPE was so robust in Experiments 1–3, we conducted Experiment 4 to implement a realistic method by which the LPE might be prevented in applied contexts. Specifically, in addition to prompting certainty judgments after each decision, we allowed participants to see the face pairs again for any trial on which they responded “not sure.” This is similar to a situation wherein a security agent or clerk is unsure whether the photo on an ID matches the cardholder. When this happens, the observer will take a second look. In Experiment 4, we sought to determine whether the LPE in unfamiliar face matching can be eliminated by allowing “second looks” on trials on which the observer is uncertain of his/her response.

Method

The procedures were the same as those in Experiment 3, except that, when participants responded “not sure” to the certainty prompt, they viewed the face-matching trial again and made a second decision.

Participants—Fifty-three Louisiana State University students participated for partial course credit. Twenty-seven ($M_{\text{age}} = 19.5$, 17 females) were randomly assigned to the HP condition, and 26 ($M_{\text{age}} = 19.8$, 19 females) were assigned to the LP condition. Thirty-eight participants self-identified as White, 8 as Black, 6 as Asian, and 1 as Hispanic. All participants reported normal or corrected-to-normal vision.

Results and discussion

Analysis on participants’ final miss rates once again revealed a reliable main effect of prevalence, $F(1, 51) = 49.71, p < .001, n^2_p = .49$. Despite being prompted to reconsider their decision after every trial, and given the opportunity to have a second look at any trials that produced uncertainty, participants in the LP group made more miss errors ($M = .44, SE = .02$) than did participants in the HP group ($M = .22, SE = .02$). As is shown in Fig. 7, this effect was no different than the effect observed when miss rates were calculated using participants’ initial responses, $F(1, 51) = 46.22, p < .001, n^2_p = .48$. This replicates the finding from Experiment 3; making a second decision, even when that decision was accompanied by a second look at the faces, did not diminish the LPE in unfamiliar face matching. As in the prior experiments, we replicated the LPE in an item analysis on mismatching pairs, thus ensuring that the effect was not a selection artifact: Mismatching face pairs produced 43 % misses in the LP condition and 22 % misses in the HP condition, $t(228) = 8.9, p < .001$. We also replicated the differences observed in false alarm rates as a function of prevalence. In final decisions, participants in the LP group made fewer false alarms ($M = .09, SE = .01$), relative to the HP group ($M = .21, SE = .01$), $F(1, 51) = 59.24, p < .001, n^2_p = .54$. The same pattern was observed in initial false alarms, $F(1, 51) = 62.28, p < .001, n^2_p = .55$.

As in Experiments 1–3, and consistent with the visual search literature, the LPE in Experiment 4 reflected a conservative criterion (c) shift in the LP group, $F(1, 51) = 88.94, p < .001, n^2_p = .64$, without a concomitant change in d' , $F(1, 51) = 0.62, p = .44$. Whereas participants in the LP group adopted a strong conservative bias ($M = .61, SE = .05$),

participants in the HP group had no bias ($M = .02$, $SE = .04$). This difference was observed whether we calculated c from initial or final decisions.

As in Experiment 3, we observed that prompting participants to reevaluate every decision yielded more frequent response switching (relative to Experiment 2, in which participants were not overtly prompted). Also as in Experiment 3, not all participants utilized the “not sure” response, preventing us from conducting full statistical analyses on the conditionalized accuracy data. Whereas 20 participants in the HP group responded “not sure” to both match and mismatch trial types, only 8 participants in the LP group used that response for both trial types. Considering overall frequencies, in the HP condition, there was a total of 6,534 trials (27 participants $\times$ 242 trials, with equal matching and mismatching trials (3,267 each). Participants responded “not sure” 111 times on matching trials (a “not sure” rate of 3.40 %), and 120 times on mismatching trials (a rate of 3.67 %). In the LP condition, there were a total of 6,292 trials (26 participants $\times$ 242 trials), with 5,668 matching trials and 624 mismatching trials. Participants responded “not sure” 68 times on matching trials (1.20 %) and 18 times on mismatching trials (2.88 %). There was thus a trend toward more corrections to mismatches in the LP condition (in the proportions of relevant trials), but corrections did not make a compelling change in the LPE. As is shown in Table 5, the “second look” manipulation served to modestly decrease the magnitude of the LPE, relative to Experiments 1–3. Whereas the miss rate in the LP group was reliably over 45 % in our previous experiments, the miss rate during “second looks” to infrequent mismatches decreased to 34 % in Experiment 4, but this estimate is based on a very small handful of observations. The more compelling point is that perceivers truly “missed” rarely occurring mismatching face pairs when viewing them the first time.

General discussion

As has been repeatedly observed (Bindemann et al., 2010; Burton, White, & McNeill, 2010; Clutterbuck & Johnston, 2002; Kemp et al., 1997; Megreya & Bindemann, 2009) and was replicated in the present research, unfamiliar face matching is surprisingly difficult. As a novel contribution, we also found that it is powerfully affected by contextual statistics. Across four experiments, we examined the effect of mismatch prevalence on observers’ ability to detect identity mismatches in unfamiliar face matching. In each experiment, we observed a robust LPE; observers were far more likely to miss identity mismatches when such occurrences were rare. This effect persisted despite allowing participants time to correct their initial responses (Experiment 2), forcing participants to evaluate their certainty of every response (Experiment 3), and allowing participants a “second look” when they were uncertain about initial decisions (Experiment 4).

Although the only previous attempt to examine the role of prevalence in unfamiliar face matching (Bindemann et al., 2010) revealed no LPE, major methodological differences may explain why our experiments elicited a strong prevalence effect and theirs did not. For example, whereas Bindemann et al. used photographs of individuals taken approximately 15 min apart, we used photographs with an average of 1.5 years in between them. Furthermore, we included 24 mismatch trials per LP participant and did not inform them of their prevalence group assignment; Bindemann et al. informed their participants of their

prevalence condition and used only 1 mismatch trial at the end of the experimental block. Therefore, whereas our procedure included feedback that could affect participants' criteria over the course of the experiment, no such learning could occur in the procedure used by Bindemann et al. Although other minor methodological differences exist, our experiments have revealed that, when conditions more closely approximate those faced by individuals who are tasked with identity verification in natural settings, the LPE is robust and difficult to alleviate.

To date, most research on unfamiliar face matching has focused on the effects of perceptual or contextual characteristics on performance. For example, compromised or time-limited viewing conditions diminish performance in face matching (e.g., Özbek & Bindemann, 2011), and changes in lighting or viewpoint also reduce performance (e.g., Hill & Bruce, 1996). Myriad visual changes can also occur within individuals, particularly as time elapses between creating an ID photograph and later ID verification. Within relatively short time periods, individuals can change in weight, hair color/style, facial expressions, facial characteristics (e.g., a beard or glasses), and other features. Although some of these changes occur over a period of years and are indeed partly why photo IDs eventually expire, researchers have documented large changes that can occur over shorter time spans, ranging from days to weeks.

In a demonstration of the pervasiveness of such within-person variability, Jenkins, White, Van Montfort, and Burton (2011) had U.K. participants sort 40 photographs into separate piles, such that each pile should only contain photographs of the same person. Unbeknownst to participants, only two individuals (both Dutch celebrities) were actually represented in the set of 40 photographs. No participants accurately sorted the photos into two piles; instead, the median number was 7.5. In contrast, almost all Dutch participants, for whom the celebrities were familiar, sorted the photographs into two piles, ruling out the possibility that there was some inherent problem in the photo selections. Bindemann and Sandford (2011) obtained a similar result: Using three different photo IDs of the same person, they found that participants' ability to match the ID to the individual in a lineup varied substantially across different IDs but was also generally poor. Only 38 % of their participants matched the ID card to the correct lineup face for all three IDs, and most people did not realize that all three cards depicted the same person. Performance improved (to 85 %) only when participants viewed all three IDs simultaneously and were explicitly told that they depicted the same person. With such large variability within photographs of the same individual, it is not surprising that researchers have focused so much empirical work studying these factors in unfamiliar face matching. Although our research design allowed for large within-person variability, our primary focus was on a less studied, nonperceptual factor in unfamiliar face matching—mismatch prevalence—which proved to be another powerful variable.

Most prior research on unfamiliar face matching has documented surprisingly poor performance under both realistic and idealized conditions (Bindemann et al., 2010; Burton et al., 2010; Clutterbuck & Johnston, 2002; Kemp et al., 1997; Megreya & Bindemann, 2009). With the exception of a few studies wherein lineups were the primary focus (and Bindemann et al., 2010), observers are typically shown equal numbers of match and mismatch trials, a proportion that is rarely observed in applied contexts. This is an important and understudied

characteristic, since target prevalence dramatically affects performance in signal detection tasks like visual search (Wolfe et al., 2005; Wolfe et al., 2007; Wolfe & Van Wert, 2010). In visual search, when observers search for relatively rare targets (e.g., weapons in luggage, tumors on radiological scans), their miss rates increase. This effect is not attributed to changes in the ability to detect targets, since observers' sensitivity (d') does not change. Instead, their thresholds for responding (their criterion, c) become higher, or more difficult to meet. Although some researchers have reported success diminishing the LPE in visual search (e.g., by giving the observers a chance to correct their initial responses; Fleck & Mitroff, 2007), others have found it to be a persistent source of error, resistant to many efforts to diminish it (Wolfe et al., 2007).

Whereas Bindemann et al. (2010) suggested that the natural variability in diagnostic features may attenuate the LPE in face matching, we suggest the opposite, for two reasons. First, empirically, even though our materials contained high within-person variability, observers were clearly sensitive to the frequency with which targets (mismatches) were encountered, and they dynamically adjusted their thresholds accordingly. With two levels of mismatch prevalence (low, 10 %, and high, 50 %), we found large context-sensitive criterion shifts. Like the LPE in visual search, observers adopted more stringent thresholds under conditions of low prevalence, increasing the evidence necessary to respond "mismatch," which increased miss rates.

Second, theoretically, by combining high within-person variation with a prevalence manipulation, our method makes very clear predictions from a signal detection perspective: In our procedure, people are often shown rather dissimilar photographs, but depicting the same individual, and other dissimilar photographs that depict different individuals (as in Fig. 1). This is a classic signal detection problem, with "signal" and "noise" distributions that naturally vary in their evidence strength and that surely overlap as many instances are sampled. Therefore, an observer must adopt a criterion that optimizes either performance or payoffs. When people made classification errors, they received negative feedback in the form of time penalties, with slightly longer timeouts for miss errors, relative to false alarms. Given 50 % target prevalence, people would be expected to balance out their misses and false alarms or to lean in favor of false alarms, as a "centered" or slightly liberal criterion would be optimal. (Had participants been sensitive to the nonsymmetric penalties we imposed, they should have adopted liberal thresholds, accepting more false alarms and minimizing misses; we did not observe this pattern.) By contrast, when mismatches occur only 10 % of the time and without the perceptual task becoming any easier, the optimal approach to minimize penalties is to report "mismatch" less often, reserving it for cases that are especially clear. Given a low base-rate and a challenging perceptual task, the LPE in face matching appears virtually inevitable.

Is it possible that, under LP conditions, observers simply make more response execution errors (i.e., responding too quickly with the same, often-used response option), despite having detected the target? After all, we did observe speed-accuracy trade-offs in our RT data: When participants rarely encountered mismatches, "match" RTs were short, regardless of accuracy. Although there is evidence that this can partly explain the LPE in visual search (Fleck & Mitroff, 2007; Rich et al., 2008), we suggest that the LPE in face matching is

largely immune to such response execution errors. Instead, the high miss rate observed in LP conditions is better explained by changes to the “quitting threshold” for mismatch responses. Consider a framework such as the diffusion model (e.g., Ratcliff, 2006; Ratcliff & Starns, 2009), wherein perceptual evidence “moves” perception toward either of two thresholds: Changes in the relative prevalence of matches and mismatches will induce asymmetric criterion shifts for their perceptual thresholds, simultaneously changing both response rates and RTs. In this framework, a speed-accuracy trade-off is expected, but it reflects perceptual and cognitive processes, rather than observers falling into a rhythmic motor routine. Indeed, Wolfe and van Wert (2010) modeled LPE data from visual search using the diffusion drift model, finding an excellent fit to both response rates and RTs. Similar to the present study, they also found that observers’ sensitivity (d') and target-present RTs were unaffected by prevalence; only criteria were affected.

In the present study, although the RT data suggested a speed-accuracy trade-off in Experiment 1, as can be seen in Fig. 8, we were able to successfully “slow down” participants in Experiments 2–4 but did not eliminate the LPE. We also did not observe changes in d' associated with changes in RTs. This provides a conceptual replication of the persistent LPE in visual search; when Wolfe et al. (2007) gave participants “speeding tickets” for responding too fast, they still observed increased misses with decreases in target prevalence. In fact, Wolfe et al. (2007) found that this speed-accuracy trade-off did not reflect quick or careless responding when targets were rare. They reasoned that, if the RT effect in the LPE was the result of careless lapses of attention, then the joint miss rate of two independent observers should be the product of their individual miss rates (e.g., if each observer misses 25 % of targets, then together, they should only miss approximately 6 %). They observed, however, that joint error rates were only slightly lower than independent error rates, ruling out an interpretation based on speeded carelessness. In future research, we plan to determine whether similar effects arise in face matching.

The present research yielded findings relevant to both theoretical and applied accounts of unfamiliar face matching. Theoretically, we observed that face matching, like visual search, is inherently a signal detection task, albeit perhaps noisier than standard, laboratory visual search. When faced with a poorly specified search task (e.g., “search each face until diagnostic matching or mismatching information is acquired”), observers are sensitive to target prevalence; we repeatedly observed inflated miss rates and fewer false alarms in conditions with infrequent mismatches. This LPE does not manifest itself as changes to observers’ perceptual capabilities: Despite being equally sensitive in low- and high-prevalence conditions, observers adopted more conservative criteria when mismatches became uncommon. In other words, they required more evidence to decide that mismatches had been identified. From an applied perspective, future research should aim to determine which facial features provide the most diagnostic information under these conditions and under what circumstances might conservative criteria be relaxed. Because unfamiliar face matching is ubiquitous in applied contexts, often with serious societal implications, understanding the cause of the LPE is a critical first step in decreasing the frequency of missed impostors.

Acknowledgments

We thank Laura Betancourt, Caroline Rausch, Stephanie Incaprera, and Tan Vu for their assistance with data collection and Rachelle Friedman for assistance with stimulus creation. Special thanks to Caroline Rausch, who tirelessly created such compelling “fake IDs.” We also thank Steve Mitroff and Markus Bindemann for comments on earlier versions of the manuscript.

Appendix Response time analyses for Experiments 2–4

As was noted in the Results section of Experiment 1, the RT results from all remaining experiments (2, 3, and 4) produced patterns virtually identical to those reported in Experiment 1. The following tables provide the observed means and standard errors (Table 5) and their corresponding ANOVA results (Table 6):

Table 5

Mean response times (in milliseconds, with standard errors in parentheses) as a function of response type and prevalence condition, Experiments 2–4

Exp	Prevalence	Hit	Miss	FA	CR
2	Low	2,608 (128)	2,135 (147)	2,823 (144)	1,742 (78)
	High	2,227 (139)	2,814 (160)	2,522 (157)	2,191 (84)
3	Low	2,648 (140)	2,166 (188)	2,996 (191)	1,749 (109)
	High	2,228 (158)	2,775 (213)	2,575 (216)	2,258 (123)
4	Low	2,476 (225)	2,798 (217)	4,029 (254)	2,275 (137)
	High	2,192 (220)	2,702 (213)	2,581 (249)	2,177 (135)

Note. FA and CR denote false alarms and correct rejections, respectively.

Table 6

Results from ANOVAs on the response time data, Experiments 2–4

Exp	Effect	df	<i>F</i>	<i>p</i>	η^2_p
2	Accuracy (A)	1, 89	61.42	<.001	.41
2	Trial type (TT)	1, 89	13.50	.001	.13
2	Resp. cond (RC)	1, 89	1.49	.22	.02
2	Prevalence (P)	1, 89	1.31	.25	.02
2	A × TT × P	1, 89	72.85	<.001	.45
2	A × TT × P × RC	1, 89	1.10	.29	.01
3	A	1, 48	48.57	<.001	.50
3	TT	1, 48	1.92	.17	.04
3	P	1, 48	0.07	.79	.00
3	A × TT × P	1, 48	44.63	<.001	.48
4	A	1, 51	79.20	<.001	.61
4	TT	1, 51	1.75	.19	.03
4	P	1, 51	7.49	.01	.13
4	A × TT × P	1, 51	67.25	<.001	.57

References

- Bindemann M, Avetisyan M, Blackwell K. Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied*. 2010; 16:378–386. [PubMed: 21198254]
- Bindemann M, Sandford A. Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*. 2011; 40:625–627. [PubMed: 21882725]
- Bruce V, Henderson Z, Greenwood K, Hancock PJB, Burton AM, Miller P. Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*. 1999; 5:339–360.
- Bruce V, Henderson Z, Newman C, Burton AM. Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*. 2001; 7:305–327.
- Burton AM, Jenkins R, Hancock PJB, White D. Robust representations for face recognition: The power of averages. *Cognitive Psychology*. 2005; 51:256–284. [PubMed: 16198327]
- Burton AM, White D, McNeill A. The Glasgow Face Matching Test. *Behavior Research Methods*. 2010; 42:286–291. [PubMed: 20160307]
- Burton AM, Wilson S, Cowan M, Bruce V. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*. 1999; 10:243–248.
- Clutterbuck R, Johnston RA. Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*. 2002; 31:985–994. [PubMed: 12269591]
- Davis JP, Valentine T. CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*. 2009; 23:482–505.
- Fleck MS, Mitroff SR. Rare targets are rarely missed in correctable search. *Visual Cognition*. 2007; 17:195–211. [PubMed: 19498955]
- Hancock PJB, Bruce V, Burton AM. Recognition of unfamiliar faces. *Trends in Cognitive Sciences*. 2000; 4:330–337. [PubMed: 10962614]
- Hill H, Bruce V. The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*. 1996; 22:986–1004. [PubMed: 8756964]
- Jenkins R, White D, Van Montfort X, Burton AM. Variability in photos of the same face. *Cognition*. 2011; 121:313–323. [PubMed: 21890124]
- Kemp R, Towell N, Pike G. When seeing should not be believing: Photographs, credit cards, and fraud. *Applied Cognitive Psychology*. 1997; 11:211–222.
- Lindsay RCL, Pozzulo JD. Sources of eyewitness identification errors. *International Journal of Law and Psychiatry*. 1999; 22:347–360. [PubMed: 10457929]
- Malpass RS, Devine PG. Eyewitness identification: Lineup instructions and the absence of the offender. *Law and Human Behavior*. 1981; 15:43–57.
- Megreya AM, Bindemann M. Revisiting the processing of internal and external features of unfamiliar faces: The headscarf effect. *Perception*. 2009; 38:1831–1848. [PubMed: 20192132]
- Megreya AM, Bindemann M, Havard C. Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica*. 2011; 137:83–89. [PubMed: 21459354]
- Megreya AM, Burton AM. Unfamiliar faces aren't faces: Evidence from a matching task. *Memory & Cognition*. 2006; 34:865–876. [PubMed: 17063917]
- Megreya AM, Burton AM. Hits and false positives in face matching: A familiarity based dissociation. *Perception & Psychophysics*. 2007; 69:1175–1184. [PubMed: 18038955]
- Megreya AM, Burton AM. Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*. 2008; 14:364–372. [PubMed: 19102619]
- Özbek M, Bindemann M. Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research*. 2011; 51:2145–2155. [PubMed: 21864559]
- Ratcliff R. Modeling response signal and response time data. *Cognitive Psychology*. 2006; 53:195–237. [PubMed: 16890214]

- Ratcliff R, Starns JJ. Modeling confidence and response time in recognition memory. *Psychological Review*. 2009; 116:59–83. [PubMed: 19159148]
- Recinto, R. 'Missing' woman unknowingly joins search for herself. 2012. Retrieved from: <http://news.yahoo.com/blogs/sideshow/missing-woman-unknowingly-joins-search-herself-165249353.html>
- Rich AN, Kunar MA, Van Wert MJ, Hidalgo-Sotelo B, Horowitz TS, Wolfe JM. Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of Vision*. 2008; 8:1–17. [PubMed: 19146299]
- Simons DJ, Levin DT. Failure to detect changes to people in a real-world interaction. *Psychonomic Bulletin and Review*. 1998; 5:644–649.
- Van Wert MJ, Horowitz TS, Wolfe JM. Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception, & Psychophysics*. 2009; 71:541–553.
- Wells GL, Olson EA. Eyewitness testimony. *Annual Review of Psychology*. 2003; 54:277–295.
- Winer, BJ. *Statistical principles in experimental design*. New York: McGraw-Hill; 1971.
- Wolfe JM, Horowitz TS, Kenner NM. Rare items are often missed in visual searches. *Nature*. 2005; 435:439–440. [PubMed: 15917795]
- Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*. 2007; 136:623–638. [PubMed: 17999575]
- Wolfe JM, Van Wert MJ. Varying target prevalence reveals two, dissociable decision criteria in visual search. *Current Biology*. 2010; 20:121–124. [PubMed: 20079642]

Fig. 1.

Sample face pairs, with original ID photos on the left (column A), more recent photos in the middle (column B), and foil photos on the right (column C). The rows are arranged according to elapsed time between photos. The matching examples in the left A–B columns are subjectively easy to appreciate as matching; those in the right A–B columns are subjectively more challenging

Fig. 2.
Sample stimulus display for a positive match trial, with our recent photograph as the larger (simulated “live”) face and the older student ID photo embedded in a simulated driver’s license

Fig. 3.
Average mismatch miss rate in the low- and high-prevalence groups in Experiment 1. Error bars represent standard errors of the means

Fig. 4.

Average response times as a function of decision type and target prevalence in Experiment 1. Error bars represent standard errors of the means

Fig. 5.

Average mismatch miss rates in the low- and high-prevalence groups in the standard and correction conditions in Experiment 2. Error bars represent standard errors of the means

Fig. 6. Average mismatch miss rates in the low- and high-prevalence groups, calculated on participants' initial and final responses in Experiment 3. Error bars represent standard errors of the means

Fig. 7. Average mismatch miss rates in the low- and high-prevalence conditions in Experiment 4, based on participants' initial and final responses. Error bars represent standard errors of the means

Fig. 8.

Average response times in low and high target prevalence conditions in each of the experiments reported here. Error bars represent standard errors of the means

Table 1
Characteristics of the depicted students, in terms of years between photographs and self-reported ethnic identities

Race/Ethnicity								
Age	Change	Asian	Black	Hispanic	Indian	Middle Eastern	White	Bracial [/] Total
0		9	2	4	2	2	15	2 37
1		4	4	9	2	1	70	3 93
2		5	2	12	2	0	41	1 63
3		3	0	5	0	0	24	4 36
4		2	0	1	0	1	7	2 13
5+		0	0	0	0	0	3	0 3
Total		23	8	30	6	4	160	12 245

[/] Bracial individuals included 6 Asian-White, 2 Hispanic-White, and 1 Asian-Hispanic.

Table 2

Accuracy of corrected responses in Experiment 2

Trial Type	Prevalence	
	Low	High
<i>Match</i>	100 % (4)	75 % (4)
<i>Mismatch</i>	n/a (0)	89 % (9)

Note. Values in parentheses represent raw frequencies of observations

Table 3

Accuracy of corrected responses in Experiment 3

Trial Type	Prevalence	
	Low	High
<i>Match</i>	62 % (149)	42 % (78)
<i>Mismatch</i>	21 % (41)	44 % (140)

Note. Values in parentheses represent raw frequencies of observations

Table 4

Accuracy of second decisions (i.e., trials on which participants responded “not sure” and viewed the face pairs again) in Experiment 4

Trial Type	Prevalence	
	Low	High
<i>Match</i>	54 % (68)	75 % (111)
<i>Mismatch</i>	66 % (18)	79 % (120)

Note. Values in parentheses represent raw frequencies of observations