Journal of Experimental Psychology: Applied

The Effect of the Proportion of Mismatching Trials and Task Orientation on the Confidence–Accuracy Relationship in Unfamiliar Face Matching

Rachel G. Stephens, Carolyn Semmler, and James D. Sauer

Online First Publication, August 14, 2017. http://dx.doi.org/10.1037/xap0000130

© 2017 American Psychological Association
1076-898X/17/$12.00 http://dx.doi.org/10.1037/xap0000130

The Effect of the Proportion of Mismatching Trials and Task Orientation on the Confidence–Accuracy Relationship in Unfamiliar Face Matching

Rachel G. Stephens
University of New South Wales and University of Adelaide

Carolyn Semmler
University of Adelaide

James D. Sauer
University of Tasmania

Unfamiliar, one-to-one face matching has been shown to be error-prone. However, it is unknown whether there is a strong relationship between confidence and accuracy in this task. If there is, then confidence could be used as an indicator of accuracy in real-world face matching settings such as border security, where the objectively correct decision is typically unknown. Two experiments examined the overall confidence–accuracy relationship, as well as the relationship for positive (match) and negative (mismatch) decisions. Furthermore, they tested whether these relationships were affected by factors relevant to applied face matching settings: the proportion of mismatching trials (PMT), and the task orientation of the decision-maker (look for matches, or look for mismatches). Both calibration analyses and signal detection methods were applied to assess performance. The results showed that confidence can have a high correspondence with accuracy overall, regardless of task orientation but with small effects of PMT. Thus, confidence is promising as an indicator of accuracy in face matching. However, PMT systematically produces large detrimental effects on the confidence–accuracy relationships for positive and negative decisions, when considered separately. Signal detection measures help with understanding these effects and proposing future research directions for improving the relationships.

Keywords: face matching, confidence and accuracy, calibration analysis, signal detection theory, base-rate effects

Supplemental materials: http://dx.doi.org/10.1037/xap0000130.supp

Identity verification using facial images is an important procedure in numerous settings, including border security, controlling access to restricted areas, criminal investigations, and selling age-restricted goods such as alcohol. While automated face recognition systems can be highly accurate (O'Toole, An, Dunlop, Natu, & Phillips, 2012; Phillips & O'Toole, 2014), many instances arise where humans are required to verify the identity of individuals from images in a one-to-one comparison, or at least have the final say on recognition output from automated systems (White, Dunn, Schmid, & Kemp, 2015). Troublingly, people can be surprisingly poor at this kind of face matching task: deciding whether two images show the same (unfamiliar) person. Error rates are around 20% on a typical test (Burton, 2013; Burton, White, & McNeill, 2010; Megreya & Burton, 2007), even for some trained and experienced passport officers (White, Kemp, Jenkins, Matheson, & Burton, 2014). Indeed, errors are also frequent when matching a live person to a photograph (Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008). This could translate to an unacceptably high number of errors in applied settings such as border security. Further, in these settings there is often no way to be sure that the correct decision has been made. However, the decision-maker's confidence in a decision may offer an indicator of accuracy. If confidence can be shown to predict accuracy in face matching, considerations of confidence could be included in procedures in applied settings.

The Confidence–Accuracy Relationship

Despite the potential theoretical and applied benefits, few studies have investigated the relationship between confidence and accuracy in face matching. However, much research has focused on confidence as an independent marker of decision accuracy in the context of *memory* for faces (e.g., Brewer & Wells, 2006; Olsson & Juslin, 2002; Weber, Woodard, & Williamson, 2013). Promisingly, this work suggests that in the absence of distorting factors (such as a delay between the decision and confidence rating) the confidence–accuracy relationship should be strong (for a review, see Brewer & Weber, 2008). Furthermore, long-standing

Rachel G. Stephens, School of Psychology, University of New South Wales and School of Psychology, University of Adelaide; Carolyn Semmler, School of Psychology, University of Adelaide; James D. Sauer, Division of Psychology, University of Tasmania.

This work was supported by a grant from the Department of Prime Minister and Cabinet, Research Funding for National Security, to the second author. The authors thank John C. Dunn for his useful suggestions and Victoria Gilliland for her assistance with the face stimuli.

Correspondence concerning this article should be addressed to Rachel G. Stephens, School of Psychology, University of New South Wales, Sydney NSW 2052, Australia. E-mail: r.stephens@unsw.edu.au

theories of judgment and decision making imply that confidence and accuracy should be related. For example, signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005) and evidence accumulation models (e.g., Horry & Brewer, 2016; Vickers, 1979) propose that strength of evidence determines both confidence and accuracy.

Existing work suggests that this strong relationship could extend to unfamiliar face matching. White, Kemp, Jenkins, and Burton (2014) reported that mean confidence tends to be higher for correct than for incorrect decisions. Bruce et al. (1999) plotted percent correct against the level of confidence (as indicated on a 1 to 5 scale) for responses to a one-to-many face matching task. When whole faces were available for comparison, the relationship between confidence and accuracy appeared quite strong. However, this task asked participants to select the face that looked most like a target face from an array of 10 faces, and participants knew the target was always present in the array. It is unknown whether a different confidence–accuracy relationship might be seen in simpler one-to-one face matching, which does not permit a strategy of identifying the most similar face.

Our goal was to further test and understand the confidence–accuracy relationship in one-to-one face matching, and address some important unexplored questions. First, is there a strong relationship between confidence and accuracy for both positive and negative decisions (i.e., a decision that "both images show the same person" vs. "the images are of different people," respectively)? In applied settings, errors in both decisions could have serious consequences, so it is useful to know how closely confidence tracks accuracy for both. Second, are these confidence–accuracy relationships affected by the task orientation of the decision-maker (searching for matches or mismatches), or the frequency of mismatching images?

The Positive–Negative Asymmetry

It is important to explore whether there is a meaningful confidence–accuracy relationship for both positive ("same person") and negative ("different people") decisions in face matching. The two relationships might not be equivalent, because people's accuracy for matching face pairs can be unrelated to their accuracy for mismatching faces (Megreya & Burton, 2007). Furthermore, related work in the domains of eyewitness identification and face recognition has shown a *positive–negative asymmetry* (e.g., Sporer, Penrod, Read, & Cutler, 1995; Weber & Brewer, 2004, 2006). The effect also extends to standard list-learning studies of memory (Mickes, Hwe, Wais, & Wixted, 2011). In short, for positive decisions ("that *is* the person/item that I saw before") there is a strong, generally linear confidence–accuracy relationship, whereas for negative decisions ("that is *not* the person/item I saw before") there is little or no evidence of a meaningful relationship (see Brewer & Wells, 2006; Weber & Brewer, 2003, 2004, 2006). Although people might be able to base their confidence for "old"/positive recognition decisions on the similarity between the test item and a specific item in memory, this simple one-to-one comparison is not available for "new"/negative decisions, which could be the cause of the difficulties (Lindsay et al., 2013; Sauerland, Sagana, & Sporer, 2012; Weber & Brewer, 2004, 2006).

Given that the face matching task involves a one-to-one comparison of two images (regardless of whether a decision is positive or negative), without having to retrieve specific faces from long-term memory, there could be no positive-negative asymmetry in this context. However, if other factors contribute (e.g., difficulties in scaling evidence of difference as opposed to evidence of match to the confidence response options), it may also appear in face matching. Furthermore, two factors that might affect the presence or extent of the asymmetry are task orientation and the proportion of mismatching trials.

Task Orientation

When considering whether the positive–negative asymmetry may occur in face matching, a complicating factor is the definition of a "positive" or a "negative" decision. In a recognition memory test, the task is to judge whether the presented stimulus matches a remembered item, and a positive decision is typically considered to be identifying a stimulus as "old" (i.e., that it matches the remembered item). However, a face matching task could involve either a goal that orients the decision-maker toward looking for a person that does match (e.g., this occurs when looking for a person of interest on a security watch list or identifying an unknown individual) *or* looking for "imposters" that do not match (e.g., watching for the presentation of a fake identity document). Thus, depending on task orientation, a positive decision could sensibly be defined as responding either "match" or "imposter," both for the decision-maker and/or for an experimenter's subsequent data analysis. Indeed, in contrast to the memory literature, some face matching research has treated nonmatch decisions as positive or "chose-target," at least for data analysis (e.g., Papesh & Goldinger, 2014).

If the asymmetry in the confidence–accuracy relationship extends to face matching, it is important to test whether this is affected by what counts as a positive decision or target for the decision-maker. In signal detection terms, from the decision-maker's perspective the target or "signal" is a match for a *match-orientation*, but would be a nonmatch for an *imposter-orientation*. It is possible that the positive-negative asymmetry observed in recognition memory tasks actually arises because of the orientation that the decision-maker uses to evaluate evidence for the decision. For example, people may overweight or focus on evidence that supports a positive decision (analogous to confirmation bias; Koriat, Lichtenstein, & Fischhoff, 1980), which is usually "old" decisions in recognition memory. This bias could lead to difficulty in differentiating levels of confidence for a negative decision. Therefore in face matching, under a match-orientation people may similarly focus on evidence that supports a match-decision (e.g., shared features) and produce the same kind of asymmetry (with no confidence–accuracy relationship for "different people" decisions). In contrast, under an imposter-orientation people may focus more on evidence for imposters (e.g., nonmatching features) and thus reduce or even reverse the asymmetry (toward no confidence–accuracy relationship for "same person" decisions instead).

We performed an important and novel test of this possibility in Experiment 1, by defining the task and response options for the participant as discriminating either "*match* versus *not a match*" or "*imposter* versus *not an imposter.*" Note that from this point onward—including in our data analysis—we objectively defined a positive decision to be a “same person” choice for all conditions (i.e., “match” or “not imposter”), but manipulated task orientation from the decision-maker’s perspective. We investigated whether this orientation affects the confidence–accuracy relationship overall (i.e., collapsed across positive and negative decisions), and also separately tested whether an imposter-orientation reduces or reverses the positive-negative asymmetry.

Proportion of Mismatching Trials (PMT)

A second factor that is likely to affect the confidence–accuracy relationship and is important for applied contexts is the proportion of mismatching trials (PMT). The majority of face matching studies use 50% mismatching trials (see Bindemann, Avetisyan, & Blackwell, 2010). However, it may be expected that PMT is significantly lower in many applied contexts. For example, in passport identity verification, the mismatch rate could be as low as 0.0075% (see Hetter & Cripps, 2014).

On one hand, changes in PMT are likely to affect accuracy. Studies in the domain of visual search have demonstrated a low-prevalence effect, such that target stimuli that are infrequent are often missed (Wolfe, Horowitz, & Kenner, 2005; Wolfe et al., 2007). Similar effects have been observed for rare nonmatches in one-to-one face matching. Across four experiments, Papesh and Goldinger (2014) found that for 10% PMT the mismatch error rates were around 45%, roughly doubling the rate found for 50% PMT (i.e., 20% errors). Furthermore, this inflated error rate persisted even when participants were given an opportunity to correct responses or could choose to view a given face-pair again and make a second decision.

On the other hand, PMT is also likely to affect confidence, and its relationship with accuracy. Indeed, in sensory discrimination tasks, experienced changes in the relative frequency of one stimulus over another have been shown to influence not only response probability, but also response latency and confidence (Van Zandt, 2000; Vickers, 1985). Importantly, even if people’s ability to discriminate matches and mismatches is high, if the prior probability of a mismatching pair is low, then the proportion correct could be quite low for “different people” decisions, yet quite high for “same person” decisions (see Parasuraman, Hancock, & Olofinboba, 1997; Szalma, Hancock, Warm, Dember, & Parsons, 2006). Put simply, when matches are very common, positive decisions are likely to be correct, and negative decisions are likely to be incorrect. Therefore, if people do not correctly adjust their decision and confidence thresholds to account for unequal proportions of matches and mismatches, overall calibration will deteriorate (cf. Ferrell & McGoey, 1980; Smith & Ferrell, 1983). Furthermore, any positive–negative asymmetry is quite likely to be more extreme, relative to when matching and mismatching pairs are equally likely. If the effect of rare mismatches on proportion correct occurs across levels of confidence for positive and negative decisions (i.e., high accuracy for “same person” decisions and low accuracy for “different people” decisions), then the confidence–accuracy relationship would reflect dramatic underconfidence for positive decisions, and overconfidence for negative decisions.

Indeed, face recognition and eyewitness identification studies support that PMT could affect the asymmetry in the confidence–accuracy relationship in this manner. Face recognition studies by Weber and Brewer (2003, 2004, 2006), though not directly comparable, show a larger asymmetry with 33% “old” test trials (Weber & Brewer, 2006) than with 50% “old” test trials (Experiments 1 and 2 of Weber & Brewer, 2004, also 2003). Further, Brewer and Wells (2006) randomly sampled different proportions of target-absent trials from their data in an eyewitness identification study and compared the asymmetry for target-absent base rates of 50%, 25%, and 15%. Although this study differs from tasks in which the observers *experience* the various PMTs, there was again a significant impact of base rate on the asymmetry. Across all of these face recognition and eyewitness studies, as the proportion of “old” (or target-present) trials increased (equivalent to PMT lowering), the confidence–accuracy relationship systematically shifted toward more extreme underconfidence for positive decisions and overconfidence for negative decisions.

To test whether these PMT effects occur in face matching, in Experiment 1 we manipulated PMT, as 20%, 50%, or 80% in a given block of face matching trials. The 20% and 80% levels were chosen to mirror each other in terms of the proportion of (subjectively defined) targets for each task orientation. Note that under a match-orientation the 20% PMT block had 80% of trials as matching face pairs, which were targets from the point of view of the decision-maker. In contrast, under an imposter-orientation the 80% PMT block had 80% of trials as mismatching face pairs, which were instead targets for the decision-maker.

Methods for Assessing Performance

In order to assess the relationship between confidence and accuracy, we used the calibration approach that has been applied in the eyewitness memory and face recognition literature (e.g., Brewer & Wells, 2006; Weber & Brewer, 2003, 2004; Weber, Woodard, & Williamson, 2013). This approach has been instrumental in demonstrating that (under certain conditions) confidence can be used as an indicator of accuracy for these memory-based face tasks. There are no clearly defined criteria for determining that confidence is an effective predictor. However, at a minimum, the level of confidence should have some correspondence with proportion correct and help to discriminate correct and incorrect decisions, offering more information about the probability that a face pair is a match than the binary judgment alone. Calibration plots and statistics help to assess this, though the statistics are primarily useful for comparing relative performance under different conditions.

A key component of the calibration approach is the calibration curve, which plots the proportion of correct responses at each level of reported confidence (usually aggregated across participants). Perfect calibration would be reflected by a curve falling on the diagonal identity line. Overconfidence is captured by the curve falling underneath and to the right of the identity line, and underconfidence is captured by the curve falling above and to the left of the identity line. The particularly useful feature of calibration curves is that one can directly read off the proportion correct for a given level of confidence. Assuming that previously observed curves can be generalized to future decisions, they can thus help to address the question: when a decision-maker is, say, 70% confident, how likely is she to be correct? This is potentially important information for applications to real-world identity verification contexts.

In addition to calibration curves, we also used calibration statistics to summarize the confidence–accuracy relationship for each condition (for details see Baranski & Petrusic, 1994; Yaniv, Yates, & Smith, 1991). The calibration statistic, C, captures deviation from perfect calibration: It is a weighted average of the squared difference between confidence (as a proportion) and proportion correct. C ranges from an optimal value of 0 (*perfect calibration*) to 1 (*worst possible calibration*), though in practice scores tend to be at the lower end (e.g., see Brewer & Wells, 2006). Next, the adjusted normalized discrimination index (ANDI; Yaniv et al., 1991) captures the extent to which confidence judgments discriminate correct from incorrect decisions, and can be interpreted as a measure of the proportion of variance explained. Accordingly, ANDI ranges from 0 (*no discrimination*) to 1 (*perfect discrimination*). Lastly, the over/underconfidence statistic (O/U) captures a responder's overall tendency to report confidence that is higher (i.e., overconfidence) or lower (i.e., underconfidence) than is warranted by accuracy. It simplifies to the signed difference between the average confidence rating (as a proportion) and the overall proportion correct. O/U ranges from −1 (*extreme underconfidence*) to 1 (*extreme overconfidence*). We calculated these statistics for each individual in each condition (unless otherwise noted). For readers unfamiliar with these calibration measures, they are further illustrated and contrasted in the supplemental materials.

Another obvious way in which confidence judgments could be used to examine face matching performance was to apply signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005). With confidence ratings, we could fit signal detection models to each condition, obtaining estimates of discriminability and response bias (see Dunn, 2010). These estimates allowed us to tease apart whether the experimental factors might be influencing people's ability to discriminate matching and mismatching face pairs, and/or people's tendency or bias toward making positive decisions. Under this approach we could also plot accuracy and confidence in a different way, by constructing receiver operating characteristic (ROC) curves (as used by e.g., O'Toole et al., 2012). ROC curves plot hit rates against false alarm rates across the levels of confidence, which captures any shifts in discriminability or response bias. Note that irrespective of task orientation, we define *hits* as correctly detecting matching face pairs, and *misses* as failing to detect such matches. Thus, *correct rejections* are correctly detecting mismatching face pairs, and *false alarms* are failing to detect such mismatches.

Analyses based on signal detection theory have not typically been considered alongside the calibration approach (but see Mickes, 2015). However, they convey different information, and recent work has shown that applications of SDT can offer useful insights when comparing performance across tasks in eyewitness memory (e.g., Mickes, Flowe, & Wixted, 2012). Therefore, we applied SDT to help understand performance in unfamiliar face matching, and also to further explore how any effects on calibration relate to bias and discriminability.

Summary

We investigated the confidence–accuracy relationship in unfamiliar, one-to-one face matching. A key goal was to test whether confidence has a strong relationship with accuracy, suggesting that confidence could be used as a marker of decision-accuracy. We examined the overall confidence–accuracy relationship, as well as the relationship for positive and negative decisions. For all these relationships, we tested the effects of task orientation (look for matches, or look for imposters) and the proportion of mismatching trials (20%, 50%, or 80%). We used the calibration approach and signal detection measures to assess performance.

Experiment 1

Method

Design. We manipulated the proportion of mismatching trials (PMT) and task-orientation, forming a 3 (PMT: 20%, 50%, 80%) × 2 (task-orientation: search for matches, search for imposters) repeated measures design. There was a block of trials for each of the six conditions.

Participants. Ninety-five psychology students at the University of Adelaide participated in exchange for course credit. Ages ranged from 18 to 55 years ($M = 20.8$, $SD = 5.6$), with 41 males. The self-identified ethnicities were Caucasian/European (46.3%), Australian (24.2%), Asian (20.0%), Middle Eastern (3.2%), and others (6.3%).

Materials. Images were sourced from the Color Facial Recognition Technology (FERET) Database (Phillips, Moon, Rizvi, & Rauss, 2000) and the Glasgow Database (Burton et al., 2010; http://www.facevar.com/downloads). The FERET database contains photographic images of 994 subjects taken on up to 15 different dates between 1993 and 1996. The images are frontal profile (portrait) and have a resolution of 512 × 768 pixels. The Glasgow database contains images of 303 identities, taken on two different digital cameras as well as still images from video sequences. Additional images were also sourced from a face database assembled by the Eyewitness Research Group at the University of Adelaide. These images are of 240 identities, containing only one frontal profile (portrait) image for each identity and have a resolution of 380 × 472 pixels.

A subset of 720 images was selected for this experiment, forming 180 matching pairs and 180 mismatching pairs (see Burton et al., 2010 for examples in black and white). Pairs were selected from distinct images of faces within the same database. Nonmatch pairs were matched to description (race, gender, hair color, eye color). The photos showed entire faces, including inner and outer facial features, from the top of the head (including hair) to the bottom of the chin. The faces were presented in color, on a white background. Some faces were repositioned or resized (with the aspect ratio maintained) on the image canvas to ensure that each face region was of a similar size, with the location of the eyes appearing within a similar region. All images were stored in JPEG format, and were presented to participants at a size of 768 pixels in height by 512 pixels in width. The face pairs were randomly allocated to conditions for each participant.

Procedure. Participants completed the face matching task individually. Software for the experiments was programmed in Matlab using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). Participants were asked to imagine that they were working as a border control officer, verifying people's passports. They were instructed they would try two strategies: looking for people who match the image in their passport (match-orientation conditions), and looking for imposters who do not match the image in their passport (imposter-orientation conditions). Participants completed six blocks of trials, with a block for each condition. There were 60 randomly ordered trials per block, with either 12, 30, or 48 mismatching trials depending on PMT condition. Task orientation was counterbalanced across participants, with the first three blocks as match-orientation then the remaining three as imposter-orientation, or vice versa. For each participant, for both sets of three blocks, the levels of PMT were presented either in the order 20%, 50% then 80%, or in reverse order. Participants were instructed about the task orientation at the beginning of each block, and accordingly responded with either "match" versus "not a match," or "imposter" versus "not an imposter." They were given a reminder every 10 trials to "keep looking for matches" (or "imposters").

Note that due to an error with the labeling of image files, one designated match pair was actually a mismatch pair. This pair was correctly accounted for as a nonmatch in the analysis of the results. However, because the image pairs were randomly allocated to conditions for each participant, each participant actually had a 1.7% higher PMT in one random condition.

During each trial, a fixation point was presented for 0.5 s, followed by a face pair presented side-by-side for 2 s. The fixed presentation duration of two seconds was selected to allow high accuracy and typical viewing times that people choose to use (Özbek & Bindemann, 2011), yet control viewing time across participants and trials. Participants reported whether the pair was a "match" or "not a match" (or "imposter" vs. "not an imposter") via labeled buttons on the screen, then rated their confidence from 50% ("guessing") to 100% ("certain"), in decile increments.

The initial task instructions emphasized accuracy and asked participants to use all of the confidence options. To help motivate participants, they were also informed they would receive immediate feedback at the end of the study. Four practice trials were completed with celebrity faces, with two trials per task-orientation. Though participants were introduced to the two task-orientations at the beginning of the experiment, they were not given any specific information about PMT or that it may vary (though the imagined border control setting might have suggested low PMT).

Results

In the following sections we examine general face matching performance, then the confidence–accuracy relationship both overall and separately for positive and negative decisions. For the analyses below, where the assumption of sphericity was violated for our *F* tests, we report the *p* value after correction using Greenhouse-Geisser epsilon, signaled by *pGG*. We also confirmed the conclusions drawn from all *F* tests with nonparametric permutation tests using the "ez" package in R (Lawrence, 2013). Holm corrections were used for all post hoc tests. Bootstrap confidence intervals (95%) were calculated for all descriptive statistics using the "ez" package, and generalized Eta-Squared and Cohen's *d* (with pooled standard deviation used as the denominator—we did not have an obvious control group for all pairwise comparisons) were used as measures of effect size.

Face matching performance. Face matching performance was assessed by comparing accuracy (proportion correct) for the initial binary judgment,[1] discriminability and response bias across conditions. Descriptive statistics and bootstrap confidence intervals are presented in Table 1 (note that to permit comparisons with other research, response times are also presented in Table 1 and discussed in the supplemental materials).[2] Analysis of variance (ANOVA) results are shown in Table 2.

First examining accuracy, there was a large and significant main effect of PMT, with accuracy decreasing as PMT increased ($p <$ .001 and Cohen's $d > 0.58$ for the three PMT post hoc comparisons, averaged across task orientation). There was also a reliable main effect of task orientation, though the overall mean difference of 0.02 was quite small (Cohen's $d = 0.21$, averaged across PMT), with slightly lower accuracy under the imposter-orientation.

To examine response bias and discriminability, data from the confidence response categories were recoded to form a single 12-point scale ranging from 100 match to 50 match ("same person"), then 50 mismatch to 100 mismatch ("different people"). We then collapsed the 12 confidence levels into six bins, producing a scale that ranged from 90–100 match to 90–100 mismatch. The empirical confidence-based ROC curve for each condition (responses aggregated across participants), is shown in Figure 1. The curves are asymmetrical (and we also confirmed that the slope of zROC <1), so it is appropriate to model them with an unequal variance signal detection model (rather than assume equal variance). We fit the model to each condition using the maximum likelihood procedure described by Dunn (2010; we recommend this tutorial article). The resulting predicted ROC curves from the fitted model are included in Figure 1 as the dashed and solid lines. We also fit the same model to each participant for each condition. The corresponding scores for the aggregate- and individual-level measures of discriminability (d_a)[3] and response bias (i.e., the centermost criterion parameter; the criterion that divides "same person" and "different people" responses) are shown in Table 1.

Examining these results, the aggregate-level ROC curves lie almost on top of each other, showing that discriminability was similar across levels of PMT and task orientation. Based on the individual-level fits, there was a reliable but very small effect of PMT on discriminability (for post hoc comparisons averaging across orientation, only 50% vs. 80% PMT was reliably different, $p = .03$, Cohen's $d = 0.29$), but no effect of task orientation, nor an interaction (see Table 2). However, PMT had a larger (though modest) effect on response bias. Though it is difficult to see in Figure 1, the triangular points for 20% PMT are shifted down toward zero, and the square points for 80% PMT are shifted up toward one. This suggests that as the proportion of mismatch pairs increased, participants set a decision threshold that was less strin-

[1] We consider proportion correct (in addition to the more informative SDT measures) to assist readers with comparing our results with other face matching studies: This is often used as a measure of performance.

[2] Response latencies were measured from the point at which participants could make a response, after the offset of the presentation of a given face pair.

[3] We used an appropriate measure of discriminability to suit the unequal variance SDT model. We estimated *d* using the procedure outlined by Dunn (2010), which is the distance between the signal and noise distributions (i.e., distributions for matching vs. mismatching face pairs, respectively), in the units of the noise standard deviation. We then calculated d_a, which is a distance measure that incorporates the standard deviations of both signal and noise distributions (Macmillan & Creelman, 2005; $d_a = d/\sqrt{(1+s^2)/2}$, where *s* is the signal standard deviation, and the noise standard deviation is set to 1).

Table 1
Performance Data for Experiment 1: The Top Set of Scores Show Means, (Standard Deviations), and [95% Bootstrap Confidence Intervals]

Measure	Match-orientation			Imposter-orientation		
	20% PMT	50% PMT	80% PMT	20% PMT	50% PMT	80% PMT
Accuracy	.86 (.07) [.85, .87]	.82 (.08) [.81, .84]	.77 (.12) [.75, .78]	.86 (.09) [.85, .87]	.81 (.08) [.80, .83]	.74 (.12) [.72, .75]
Response latency (sec)	.77 (.26) [.72, .80]	.77 (.20) [.74, .80]	.81 (.25) [.77, .85]	.85 (.28) [.81, .90]	.92 (.53) [.84, .98]	.96 (.35) [.91, 1.01]
Hit rate	.89 (.09) [.87, .90]	.90 (.10) [.88, .91]	.92 (.09) [.91, .94]	.88 (.11) [.86, .90]	.91 (.08) [.90, .92]	.92 (.10) [.90, .93]
False alarm rate	.22 (.19) [.20, .25]	.25 (.15) [.23, .27]	.27 (.14) [.25, .29]	.22 (.18) [.20, .25]	.28 (.16) [.26, .30]	.31 (.14) [.29, .33]
Discriminability (d_a)—individual	2.43 (.81) [2.30, 2.56]	2.34 (1.02) [2.19, 2.49]	2.53 (1.11) [2.36, 2.69]	2.47 (1.02) [2.31, 2.62]	2.25 (.75) [2.14, 2.36]	2.58 (1.41) [2.36, 2.79]
Centermost response criterion—individual	1.09 (1.00) [.94, 1.25]	.81 (.65) [.72, .91]	.71 (.47) [.63, .78]	.99 (.66) [.88, 1.09]	.69 (.50) [.61, .76]	.57 (.44) [.51, .64]
Discriminability (d_a)—aggregate	1.84	1.82	1.89	1.82	1.80	1.79
Centermost response criterion—aggregate	.67	.62	.60	.67	.53	.49

Note. The discriminability and criterion measures were estimated from fitting an unequal variance signal detection model to each participant and condition, and to responses aggregated across participants for each condition.

gent, responding with "same person" more often—relative to the proportion of matches. That is, while the average percentage of mismatch decisions was (sensibly) lowest for 20% PMT and highest for 80% PMT (e.g., for match-orientation, mean mismatch decisions were 24.9%, 42.8%, and 59.9% for 20%, 50%, and 80% PMT, respectively), the hit rates and false alarm rates increased with PMT. In other words, participants made too few match decisions when matches were common, but too many when matches were rare. Based on the individual-level model fits, this effect of PMT on response bias was significant (for all post hoc comparisons averaged across orientation, $p \leq .001$, Cohen's d was 0.49 for 20%–50% PMT; 0.25 for 50%–80% PMT; 0.72 for 20%–80% PMT). Task orientation also had a reliable but very small effect on response bias, with a slightly more stringent (higher) criteria for the match-orientation than for the imposter-orientation (Cohen's $d = 0.23$, averaged across PMT). Again, there was no interaction effect.

In summary, increasing the proportion of mismatching face pairs reduced accuracy, which appeared to be related to more lenient decision criteria, rather than people's ability to discriminate matching and mismatching pairs. In contrast, task orientation had limited effects on any of these performance measures.

Overall confidence–accuracy relationship. Overall, there was a strong linear relationship between confidence and accuracy in each condition (see Figure 2), with calibration curves for both match-orientation and imposter-orientation conditions roughly tracking the identity line of perfect calibration. However, accuracy generally decreased (relative to confidence) as PMT increased, producing a shift from slight underconfidence toward overconfidence. The calibration statistics are summarized in Table 3, with the ANOVA results for the effects of PMT and task orientation shown in Table 4. PMT significantly affected scores on C, ANDI, and O/U. As PMT increased, there was (a) a reduction in calibration and resolution (C scores increased and ANDI decreased; i.e., respectively, there was a reduction in how well confidence tracked accuracy and how well confidence discriminated correct from incorrect decisions); and (b) a shift from overall slight underconfidence to slight overconfidence. All PMT post hoc tests (averaging across task orientation) were significant ($p < .02$), and the difference in C, ANDI, and O/U between 20% PMT and 80% PMT corresponded to effect sizes of Cohen's $d = 0.64$, 0.86, and 0.62, respectively. In contrast to PMT, the only statistically significant calibration effect for task orientation was on O/U, though this effect was very small (Cohen's $d = 0.16$, averaged across PMT). In short, PMT (but not task orientation) systematically affected overall calibration, which was nevertheless quite good in all conditions. Importantly (i.e., for applied face matching settings), calibration and resolution were best in the 20% PMT conditions, despite minor underconfidence.

Confidence–accuracy relationship for positive and negative decisions. Despite the strong confidence–accuracy relationship across all decisions, we found asymmetries when we examined positive and negative decisions separately. For data-analysis we defined positive decisions as "same person" responses, and nega-

Table 2
ANOVA Results for the Performance Measures in Experiment 1

Effect	Accuracy			Discriminability (d_a)			Centermost response criterion		
	F	p	η^2	F	p	η^2	F	p	η^2
PMT	**71.29**	**<.001^**	**.183**	**4.06**	**.022^**	**.010**	**38.9**	**<.001^**	**.064**
Task orientation	**8.17**	**.005**	**.006**	.00	.994	.000	**5.77**	**.018**	**.009**
PMT * Orientation	2.30	.103	.004	.42	.645^	.001	.07	.885^	.000

Note. The degrees of freedom were (1,94) for the main effect of task orientation, and (2,188) for the main effect of proportion of mismatch trials (PMT) and the interaction effect. Significant effects ($p < .05$) are signaled by bold font.
^ The assumption of sphericity was violated, so the Greenhouse-Geisser epsilon correction was used.

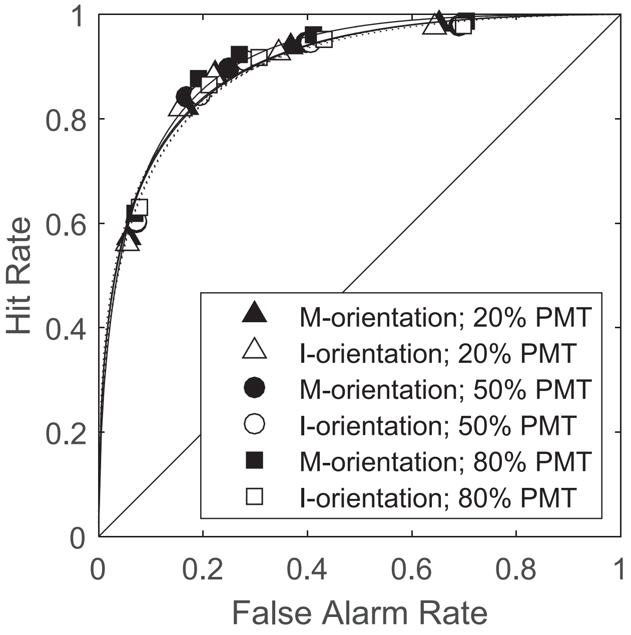

Figure 1. The points show empirical ROC curves from responses aggregated across participants in Experiment 1. "M-orientation" and "I-orientation" refer to whether participants were instructed to search for matches or imposters, respectively, and the proportion of mismatch trials (PMT) was either 20%, 50%, or 80%. Also included are theoretical ROC curves from fitting an unequal variance signal detection model to the aggregated rates: The solid lines are for match-orientation conditions, and the dashed lines are for imposter-orientation conditions.

tive decisions as "different people" responses, regardless of task orientation. The calibration curves in Figure 3 indicate that a positive-negative asymmetry was present for 50% PMT, which then changed systematically with PMT but not with task orientation. The calibration statistics are summarized in Table 5. ANDI scores are presented only for data aggregated across participants, because within each condition ANDI could not be calculated for up to 42 individuals (i.e., 44% of participants), when there were no incorrect responses. The ANOVA results for the effects of PMT and task orientation on C and O/U are shown in Table 6.

Together, Figure 3, Table 5 and Table 6 show that when positive and negative decisions are considered separately, task orientation did not affect the confidence–accuracy relationships, but PMT did. Specifically, for *positive* decisions, as PMT increased, resolution improved but calibration was reduced (C and aggregate-level ANDI scores increased), and there was a shift from under- to overconfidence. All PMT post hoc tests (averaging across task orientation) for C and O/U were significant ($p < .001$), with large effect sizes (Cohen's $d > 0.93$). *Negative* decisions showed the opposite pattern across PMT levels. The corresponding PMT post hoc tests were also all significant ($p < .005$), and all had quite large effect sizes (Cohen's $d > 0.73$) apart from C scores between 50% and 80% PMT (Cohen's $d = 0.23$).

Crucially, contrary to the general asymmetry findings in the face recognition and eyewitness memory literature, for 20% PMT the confidence–accuracy relationship actually appeared to be stronger for negative decisions than for positive decisions—at least at the aggregate level. This was captured by the ANDI scores from aggregated responses in Table 5, and the calibration curves with triangular markers in Figure 3. Unlike positive decisions, the curves for negative decisions ran roughly in parallel with the identity line, suggesting that confidence tracked accuracy quite well, though there was consistent overconfidence of around 15% to 20%.

Before concluding that task orientation had no effect on the positive–negative asymmetry, we checked for differences between the two subgroups of participants that completed the match-orientation conditions first versus the imposter-orientation conditions first. There might have been order effects, but instead the confidence–accuracy relationships were similar for both subgroups (see the supplemental materials for the calibration curves). Additionally, analyses of C and O/U scores did not indicate any important effects (beyond PMT) when we considered task orientation, PMT and subgroup as factors ($\eta^2 \leq 0.01$ for any reliable interaction effects involving subgroup).

Discussion

In our face matching task, we observed accuracy levels of 74%–86%, which is consistent with the error-prone performance observed in previous studies (e.g., Burton et al., 2010; Megreya & Burton, 2007). Task orientation had limited effects on accuracy, but increasing the proportion of mismatching face pairs decreased accuracy. Despite the error-prone performance, subjective estimates of confidence might be useful as indicators of accuracy, at least for overall performance. As expected, the overall confidence–accuracy relationship in face matching was quite strong, extending the findings of Bruce et al. (1999) to a one-to-one face matching task. Task orientation (i.e., what counted as a positive decision or target for the participant) had little effect, though there were some small but systematic effects of PMT: As PMT increased, there was a reduction in calibration and resolution, and a shift from overall slight underconfidence to slight overconfidence. The signal detec-

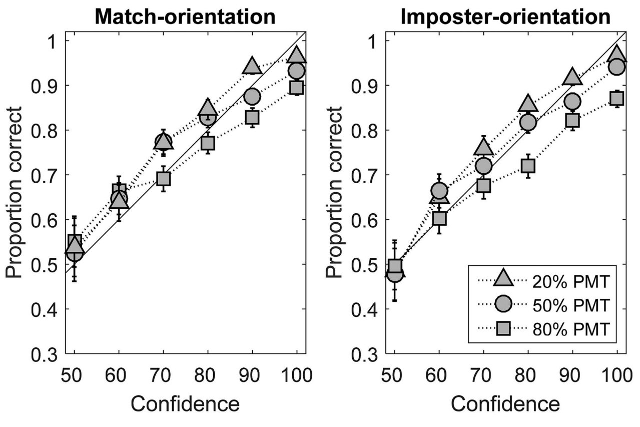

Figure 2. Experiment 1 confidence–accuracy calibration curves for each condition, collapsed across positive and negative decisions. Proportion correct is accuracy of all responses. Match-orientation conditions are in the left panel and imposter-orientation conditions are in the right panel; the proportion of mismatch trials (PMT) was either 20%, 50%, or 80%. Error bars show 95% confidence intervals for the proportions. The solid identity line marks perfect calibration.

Table 3
Mean Scores on Overall Calibration Statistics for Experiment 1 With (Standard Deviations) and [95% Bootstrap Confidence Intervals]

	Match-orientation			Imposter-orientation		
Statistic	20% PMT	50% PMT	80% PMT	20% PMT	50% PMT	80% PMT
C	.023 (.017) [.021, .026]	.028 (.019) [.025, .031]	.037 (.037) [.031, .043]	.025 (.017) [.023, .028]	.029 (.020) [.026, .032]	.042 (.037) [.037, 048]
ANDI	.237 (.178) [.209, .262]	.153 (.131) [.133, .174]	.116 (.116) [.099, .134]	.223 (.180) [.195, .251]	.182 (.152) [.158, .204]	.133 (.128) [.114, .154]
O/U	−.020 (.091) [−.035, −.006]	.001 (.113) [−.016, .018]	.033 (.138) [.010, .053]	−.015 (.098) [−.030, −.000]	.013 (.113) [−.005, .031]	.066 (.138) [.044, .086]

tion analyses suggest that this effect was related to a shift in response bias rather than discriminability. As mismatches increased, people became more biased to report "match" (relative to PMT), which aligns with previous research in which people were unaware of a change in the relative frequency of stimuli (Vickers, 1985). The good news for applied face matching settings with low rates of mismatches is that calibration and resolution were best in the 20% PMT conditions, despite some underconfidence.

More strikingly, we observed a positive–negative asymmetry in the relationship between confidence and accuracy. This asymmetry was observed for 50% PMT then varied systematically with low or high PMT, but was not affected by task orientation. For positive ("same person") decisions, as PMT increased, resolution improved (i.e., confidence discriminated correct from incorrect decisions more precisely, at least at the aggregate level) but calibration was reduced (i.e., confidence tracked accuracy less well). Most notably, there was also a dramatic shift from underconfidence to overconfidence. Negative ("different people") decisions generally showed the opposite pattern across PMT levels. Importantly, for 20% PMT the confidence–accuracy relationship (at least at the aggregate-level) was actually stronger for negative judgments than for positive. Therefore, it is not the case that the confidence–accuracy relationship is simply strong for positive decisions and weak for negative decisions. Rather, the relationships vary with the proportion of mismatching trials.

Our observed calibration curves for positive and negative decisions under the typical 50% PMT are similar to those observed in the eyewitness memory study of Brewer and Wells (2006; in particular, for the thief lineup under sensible conditions; i.e., high-similarity foils and unbiased instructions that the target may be absent in the lineup). These curves indicate some overconfidence for positive decisions and underconfidence for negative decisions, but that confidence and accuracy are somewhat related for both decisions. However, these curves are different to those observed in other face recognition studies, in which for positive decisions there is a strong confidence–accuracy relationship, but for negative decisions the curve is quite flat (Weber & Brewer, 2004) or reflects a weak relationship with high *over*confidence (Weber & Brewer, 2003). Therefore, the exact nature of the asymmetry varies across tasks. Nevertheless, our mirrored-shifts in calibration curves across different PMT levels align with the effects that have been indirectly observed in face recognition and eyewitness studies (Brewer & Wells, 2006; Weber & Brewer, 2004 vs. Weber & Brewer, 2006). Therefore, the positive-negative asymmetry is affected by the proportion of targets in a consistent manner across both memory-based face recognition tasks and the more perceptual-based face matching task. We more closely examine the reason for these effects in the General Discussion, after further testing them in Experiment 2.

Experiment 2

Two key outstanding issues remain from Experiment 1. First, the proportion of mismatches does not wholly account for the positive–negative asymmetry in face matching, because it was observed for 50% PMT. We were concerned that this may have been related to the difficulty of matching versus mismatching face pairs: The matching pairs were much easier on average (mean accuracy of 89.5% vs. 72.8%; $t(358) = 12.22$, $p < .001$). Though of course accuracy for each could have been affected by participants' decision criteria, it was possible that the correct answer tended to be more obvious for matches than for mismatches, if many matching faces were very similar but few mismatching faces appeared exceptionally dissimilar. This would be a confound and so in case it is important, in the second experiment we further examined calibration under 50% PMT, while carefully controlling the difficulty of the face pairs. We selected equivalent sets of matching and mismatching pairs, based on performance in Experiment 1. This might reduce the positive–negative asymmetry for 50% PMT.

Second, we needed to examine accuracy and confidence when the rate of mismatching trials was even lower. PMT as high as 20%

Table 4
ANOVA Results for the Overall Calibration Statistics in Experiment 1

	C			ANDI			O/U		
Effect	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
PMT	**16.66**	**<.001^**	**.060**	**25.32**	**<.001^**	**.078**	**36.33**	**<.001^**	**.054**
Task orientation	2.35	.13	.003	.96	.33	.001	**7.25**	**.01**	**.005**
PMT * Orientation	.65	.51^	.001	1.51	.22	.004	2.40	.10^	.002

Note. The degrees of freedom were (1,94) for the main effect of task orientation, and (2,188) for the main effect of proportion of mismatch trials (PMT) and the interaction effect. Significant effects ($p < .05$) are signaled by bold font.
^ The assumption of sphericity was violated, so the Greenhouse-Geisser epsilon correction was used.

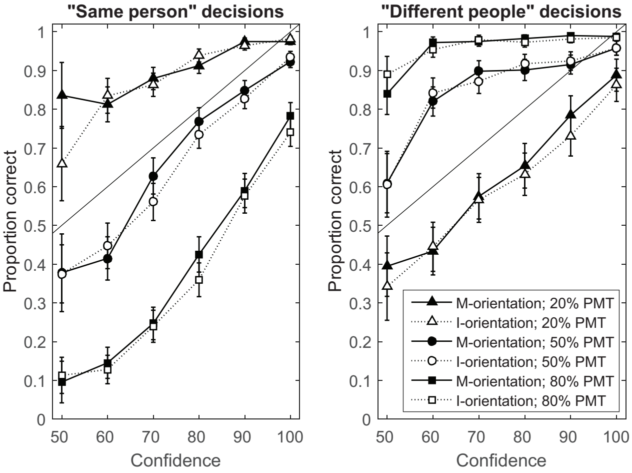

Figure 3. Experiment 1 confidence–accuracy calibration curves for positive ("same person") and negative ("different people") decisions (left and right panels, respectively), for each condition. "M-orientation" and "I-orientation" refer to whether participants were instructed to search for matches or imposters, respectively, and the proportion of mismatch trials (PMT) was either 20%, 50%, or 80%. Error bars show 95% confidence intervals for the proportions.

might be unrealistic for some applied settings such as border control. If anything, Experiment 1 showed a small reduction in false alarms for 20% mismatches (relative to 50% mismatches), but a large positive-negative asymmetry. It was unclear what to expect for even rarer mismatches. Bindemann, Avetisyan, and Blackwell (2010) found that a much lower, 2% imposter rate did not impair accuracy for detecting imposters: False alarms were below 10%, which was equal to or lower than false alarms for a 50% imposter rate (depending on condition order). In contrast, Papesh and Goldinger (2014) found that a 10% imposter rate had a large detriment when compared with a 50% imposters condition, with false alarms doubling to around 45%. Thus accuracy and its relationship with confidence may be reduced by a much lower PMT. These studies and our experiment differed in several potentially critical ways, including whether information was given about the expected PMT prior to the task, and the inclusion of penalties (pauses in the task) for incorrect responses that were harsher for false alarms than misses (as we define them). To test a more extreme PMT in our procedure, in Experiment 2 we lowered PMT to 3%. Based on Experiment 1, we might still expect a strong confidence–accuracy relationship overall. However, if the positive-negative asymmetry is systematically affected by PMT, then at 3% PMT we should see extreme underconfidence for positive judgments, and extreme overconfidence for negative judgments.

Experiment 1 found that task orientation had limited effects on the relationships between confidence and accuracy. This suggests that in face matching, people assess (or report) their confidence in much the same way, regardless of what counts as a positive judgment for them. We therefore did not include the imposter-orientation in Experiment 2.

Method

Design. This experiment manipulated PMT only (3% vs. 50%), in a repeated measures design. The match-orientation was used for both conditions. The face pairs for Experiment 2 were selected on the basis of accuracy in Experiment 1, to equalize the difficulty for matching and mismatching pairs.

Participants. The participants were 60 psychology students at the University of Adelaide, who received course credit. Ages ranged from 17 to 47 years ($M = 19.5$, $SD = 4.2$), with 16 males. The self-identified ethnicities were Caucasian (56.7%), Asian (33.3%), Middle Eastern (3.3%), and others (6.7%).

Materials and procedure. A subset of 88 matching and 32 mismatching face pairs from Experiment 1 was selected to equate difficulty. Based on performance in Experiment 1, accuracy for the matching pairs was 82.3% ($SD = 9.9$, ranging from 51.6% to 92.6%), and the accuracy for nonmatching pairs was 81.0% ($SD = 11.0$, ranging from 51.6% to 94.7%), which was a nonsignificant difference, $t(118) = 0.62$, $p = .54$.

Following Bindemann et al. (2010), for the 3% PMT condition we set the mismatching face pairs to appear at fixed trial positions within the block of 60 trials. For all participants, these mismatches appeared at trial numbers 50 and 60, to help preserve a mismatch frequency of 3%. Two face pairs of midrange difficulty were selected from the subset of 32 mismatching pairs, and were always used in the 3% PMT condition. The two pairs were Caucasian, with one pair being females and the other males. Experiment 1

Table 5
Mean Scores on Calibration Statistics for Positive and Negative Decisions in Experiment 1 With (Standard Deviations) and [95% Bootstrap Confidence Intervals]

	Match-orientation			Imposter-orientation		
Statistic	20% PMT	50% PMT	80% PMT	20% PMT	50% PMT	80% PMT
			"Same person" decisions:			
C	.032 (.023) [.028, .036]	.052 (.033) [.047, .057]	.178 (.097) [.163, .193]	.030 (.022) [.027, .033]	.057 (.038) [.051, .063]	.195 (.098) [.179, .210]
ANDI	.044	.162	.223	.067	.170	.217
O/U	−.079 (.104) [−.094, −.063]	.057 (.141) [.035, .077]	.309 (.176) [.282, .336]	−.080 (.100) [−.096, −.065]	.069 (.139) [.048, .090]	.340 (.166) [.316, .363]
			"Different people" decisions:			
C	.087 (.060) [.078, .096]	.049 (.035) [.044, .055]	.055 (.034) [.050, .060]	.104 (.073) [.093, .116]	.048 (.031) [.043, .053]	.055 (.035) [.050, .060]
ANDI	.118	.063	.041	.099	.061	.015
O/U	.074 (.190) [.045, .101]	−.108 (.121) [−.126, −.089]	−.182 (.092) [−.197, −.168]	.089 (.212) [.056, .120]	−.103 (.120) [−.121, −.086]	−.172 (.090) [−.186, −.158]

Note. ANDI scores are for data aggregated across participants, due to insufficient incorrect responses for individual-level calculations.

Table 6
ANOVA Results for Calibration Statistics in Experiment 1 for Positive and Negative Decisions

	C			O/U		
Effect	F	p	η^2	F	p	η^2
	"Same person" decisions:					
PMT	**243.08**	**<.001^**	**.557**	**1046.97**	**<.001^**	**.589**
Task orientation	3.77	.06	.003	2.71	.10	.002
PMT * Orientation	3.05	.07^	.004	3.05	.06^	.002
	"Different people" decisions:					
PMT	**42.28**	**<.001^**	**.165**	**235.51**	**<.001^**	**.362**
Task orientation	2.75	.10	.003	1.28	.26	.001
PMT * Orientation	3.21	.06^	.008	.16	.78^	.000

Note. The degrees of freedom were (1,94) for the main effect of task orientation, and (2,188) for the main effect of proportion of mismatch trials (PMT) and the interaction effect. Significant effects ($p < .05$) are signaled by bold font.
^ The assumption of sphericity was violated, so the Greenhouse-Geisser epsilon correction was used.

accuracy for each was 74.7% and 75.8%, respectively. These two pairs were randomly allocated to Trials 50 and 60 for each participant.

Participants completed two blocks in counterbalanced order, with a block for each condition (i.e., 30 people completed the 50% PMT block first). Participants were no longer given a reminder of the task orientation every 10 trials (i.e., to "keep looking for matches"). The materials and procedure were otherwise the same as the match-orientation conditions in Experiment 1.

Results

As with Experiment 1, we examine general face matching performance, then the confidence–accuracy relationship overall, and separately for positive and negative decisions. In addition to comparing the 3% and 50% PMT conditions in Experiment 2, we also make comparisons across experiments for 50% PMT (match-orientation only from Experiment 1). The latter should be interpreted cautiously, because they lack full random allocation. Nevertheless, participants in both experiments were drawn from equivalent student groups, so there is no reason for any relevant differences between them. Thus, we can test whether the positive–negative asymmetry for 50% PMT is reduced when difficulty is equated for matching and mismatching face pairs.

Face matching performance. General face matching performance in Experiment 2 is summarized in Table 7. The initial binary judgments were less accurate overall for 3% PMT than for 50% PMT, $t(59) = 4.82$, $p < .001$, $d = 0.66$. Though the false alarm rate was lower for 3% PMT, the hit rate was also lower. For 50% PMT, overall accuracy was similar across both experiments, $t(133.99) = 0.22$, $p = .83$, $d = 0.04$.

Applying the unequal variance signal detection model, we found that the differences in accuracy for 3% PMT primarily reflected shifts in response bias. The empirical and fitted ROC curves for responses aggregated across participants are shown in Figure 4 (following the same procedure as in Experiment 1), with the corresponding aggregate-level discriminability and response bias measures from the models in Table 7 (final two rows). The curves sit reasonably close together, though aggregate-level discriminability was a little lower for 3% PMT. For 50% PMT, discriminability was the same across experiments ($d_a = 1.82$). In Experiment 2, PMT appeared to affect aggregate-level response bias: The triangular points for 3% PMT (see Figure 4) are shifted further down toward zero, and the centermost criterion parameter was higher for 3% PMT than for 50% PMT, which was higher than the criteria from Experiment 1. This extends the pattern we saw in Experiment 1: Under a very low PMT, participants set more-stringent criteria, responding with "different people" more often, relative to PMT. Thus, in the condition with only 3% mismatches, participants made an average of 27.4% mismatch decisions (vs. 48.7% for 50% PMT).

For completeness, in Table 7 we also present the mean discriminability and response bias measures from individual-level model fitting for both conditions. However, discriminability estimates are probably quite inflated for 3% PMT,[4] which makes comparison with 50% PMT difficult. Nevertheless, for 50% PMT across experiments, d_a was not significantly different, $t(96.91) = 0.07$, $p = .95$, $d = 0.01$, though the centermost criterion was somewhat higher in Experiment 2, $t(86.13) = 2.13$, $p = .04$, $d = 0.37$.

In summary, performance appeared to have a U-shaped function across PMT levels from both experiments: accuracy was lowest for both very low PMT (3% PMT in Experiment 2) and higher PMT (80% PMT in Experiment 1). This appeared to be mostly related to changes in response bias, with the lowest decision criterion (and highest false alarm rate) for 80% PMT and the highest criterion (and lowest false alarm rate) for 3% PMT. For 50% PMT, overall accuracy and discriminability were not affected by the equated difficulty of matches and mismatches in Experiment 2, though there was a small difference in response bias.

Overall confidence–accuracy relationship. Experiment 2 replicated the effect that overall, there was a strong linear relationship between confidence and accuracy, extending to a very low PMT. The calibration curves in Figure 5 track the identity line of perfect calibration, though there is a tendency toward slight overconfidence for 3% PMT. The individual-level calibration statistics are summarized in Table 8. They show that although calibration and resolution were best for 20% PMT in Experiment 1, this does not extend to the more extreme 3% PMT. The confidence–accuracy relationship was somewhat less strong for 3% PMT than for 50% PMT: For 3% PMT, C scores were higher, $t(59) = 3.51$, $p < .001$, $d = 0.62$, as were O/U scores, $t(59) = 3.71$, $p < .001$, $d = 0.51$. However, ANDI scores were not significantly different, $t(59) = 0.70$, $p = .48$, $d = 0.10$.

The overall confidence–accuracy relationship was similar for 50% PMT (match-orientation) across both experiments, though in Experiment 2 the calibration curve perhaps followed the identity line even more closely. ANDI and O/U scores were not significantly different: for ANDI, $t(115.84) = 0.45$, $p = .66$, $d = 0.08$, and for O/U, $t(152.43) = 0.36$, $p = .72$, $d = 0.06$. However, C was

[4] With only two mismatching face pairs in the 3% PMT condition, false alarm rates across the ROC curves could only take three possible values (0, 0.5, and 1). Therefore, for many participants the ROC curve hugged the top-left corner, suggesting perfect discrimination. This is probably an overestimate of performance.

Table 7
Performance Data for Experiment 2: Means, (Standard Deviations), and [95% Bootstrap Confidence Intervals]

Measure	3% PMT	50% PMT
Accuracy	.75 (.14) [.72, .78]	.83 (.08) [.81, .84]
Response latency (sec)	.87 (.26) [.83, .92]	.90 (.45) [.82, .97]
Hit rate	.75 (.15) [.72, .78]	.84 (.12) [.82, .86]
False alarm rate	.12 (.25) [.06, .16]	.19 (.12) [.16, .21]
Discriminability (d_a)—individual	3.37 (2.12) [2.92, 3.78]	2.33 (1.43) [2.05, 2.57]
Centermost response criterion—individual	2.93 (2.31) [2.50, 3.40]	1.15 (1.09) [.93, 1.32]
Discriminability (d_a)—aggregate	1.68	1.82
Centermost response criterion—aggregate	1.01	.86

Note. Discriminability and criterion measures were estimated from fitting an unequal variance signal detection model to each condition, for each participant and for aggregate data.

somewhat improved in Experiment 2, $t(152.9) = 3.04$, $p = .003$, $d = 0.48$.

In summary, the overall confidence–accuracy relationship remained strong even for the more extreme 3% PMT. Nevertheless, there was a small U-shaped function across PMT levels, mirroring the effect on accuracy. Specifically, the confidence–accuracy relationship was slightly weaker (and overconfidence a little higher) for both very low PMT (3% PMT in Experiment 2) and higher PMT (80% PMT in Experiment 1). For 50% PMT, equating the difficulty of matches and mismatches in Experiment 2 made a small improvement to (the already impressive) overall calibration.

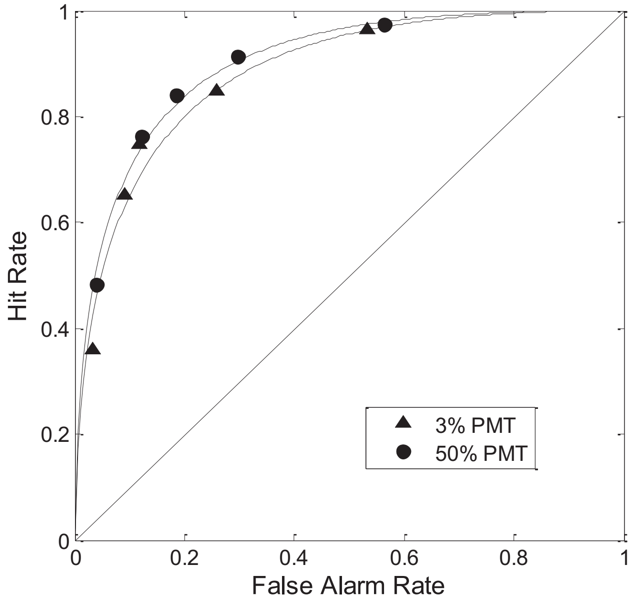

Figure 4. The points show empirical ROC curves from responses aggregated across participants in Experiment 2. Participants were instructed to search for matches, with the proportion of mismatch trials (PMT) as either 3% or 50%. Also included are theoretical ROC curves from fitting an unequal variance signal detection model to the aggregated rates.

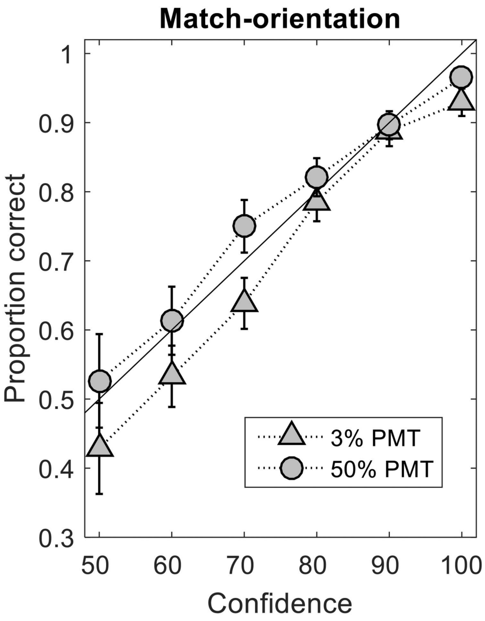

Figure 5. Experiment 2 confidence–accuracy calibration curves for each condition, collapsed across positive and negative decisions. Proportion correct is accuracy of all responses. The proportion of mismatch trials (PMT) was either 3% or 50%. Error bars show 95% confidence intervals for the proportions.

Confidence–accuracy relationship for positive and negative decisions. Despite the strong confidence–accuracy calibration across all decisions, an extreme positive–negative asymmetry was observed for 3% PMT. There was extreme underconfidence for positive decisions and extreme overconfidence for negative decisions, and this asymmetry was much more pronounced than for 50% PMT, where there was excellent calibration for both decision types (see Figure 6). Recall that in Experiment 1 we found that for 20% PMT, the confidence–accuracy relationship actually appeared stronger for negative judgments than for positive (see Figure 3). For 3% PMT the calibration curves were pushed out further toward the extremes. Confidence did not track accuracy well for either type of decision. The relationship appeared worse for positive decisions, with the calibration curve as an uninformative horizontal line, capturing the very high accuracy regardless of confidence. For negative decisions, at least confidence over 70 was predictive of higher accuracy.

Table 8
Mean Scores on Overall Calibration Statistics for Experiment 2 With (Standard Deviations) and [95% Bootstrap Confidence Intervals]

Statistic	3% PMT	50% PMT
C	.035 (.031) [.029, .040]	.020 (.012) [.018, .022]
ANDI	.179 (.163) [.146, .211]	.163 (.145) [.136, .191]
O/U	.043 (.107) [.021, .062]	−.004 (.076) [−.020, .009]

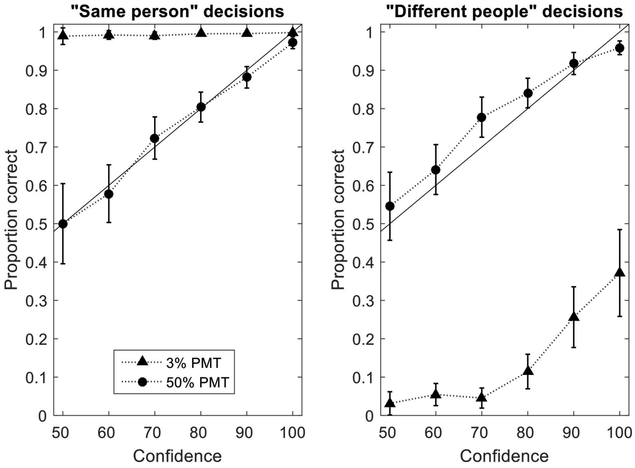

Figure 6. Experiment 2 confidence–accuracy calibration curves for positive ("same person") and negative ("different people") decisions (left and right panels, respectively), for each condition. Participants were instructed to search for matches, with the proportion of mismatch trials (PMT) as either 3% or 50%. Error bars show 95% confidence intervals for the proportions.

The calibration statistics are summarized in Table 9. As with Experiment 1, ANDI scores are presented only for data aggregated across participants. Not surprisingly, for both positive and negative decisions, confidence tracked accuracy less well under 3% PMT than under 50% PMT: C was higher (positive decisions: $t(59) = 4.05, p < .001, d = 0.69$; negative decisions: $t(59) = 19.80, p < .001, d = 3.62$), O/U was more extreme (positive decisions: $t(59) = 13.03, p < .001, d = 1.87$; negative decisions: $t(59) = 26.75, p < .001, d = 3.80$), and at the aggregate level, ANDI was lower—indeed, zero for positive decisions.

Lastly, with difficulty equated for the matching and mismatching face pairs, the confidence–accuracy asymmetry was reduced for 50% PMT in Experiment 2. Compared with 50% PMT (match-orientation) in Experiment 1, C was somewhat reduced (positive decisions: $t(152.87) = 3.91, p < .001, d = 0.61$; negative decisions: $t(152.90) = 2.21, p = .03, d = 0.35$), and O/U was closer to zero (positive decisions: $t(146.05) = 2.95, p = .004, d = 0.47$; negative decisions: $t(132.76) = 3.91, p < .001, d = 0.64$). However, aggregate-level ANDI improved only for negative decisions, where there was perhaps more room for improvement. It was possible that the reduction in the asymmetry for 50% PMT across experiments was due to participants also experiencing the very low PMT in Experiment 2, rather than being due to the equated matching and mismatching faces. However, this does not appear to be the case: calibration curves for people who completed the 50% PMT condition first show a similar reduced asymmetry (figure available in the supplemental materials).

In summary, extending the PMT effect in Experiment 1, for 3% PMT there was an extreme positive–negative asymmetry, with very underconfident positive judgments, and very overconfident negative judgments. For 50% PMT, with the equated difficulty of matching and mismatching face pairs, there was almost no asymmetry.

Discussion

Experiment 2 addressed two primary aims. First, it examined whether the positive–negative asymmetry for 50% PMT in Experiment 1 was related to differences in difficulty between matching and mismatching face pairs. Experiment 2 equated the difficulty and found almost no asymmetry for 50% PMT. Therefore, in contrast to findings in the eyewitness identification and face recognition literature, confidence can be a reliable indicator of accuracy for both positive and negative face matching decisions—at least when PMT is 50%.

Second, Experiment 2 investigated performance and the confidence–accuracy relationships under an extremely low proportion of mismatching trials (3%). For our 3% and 50% PMT conditions, we found comparable hit rates and false alarm rates to that of Bindemann et al. (2010) for 2% and 50% PMT (in their Experiment 1 at least), though our false alarms were a little higher. We too found that a very low imposter rate did not impair accuracy for detecting the rare imposters: Instead, false alarms were lower than for a 50% imposter rate, though the hit rate was also lower, reducing overall accuracy. This could be accounted for by the stricter decision criterion for deeming that a face pair was a match, and thus participants were making "mismatch" decisions too often relative to the very low PMT.

These findings contrast with those of Papesh and Goldinger (2014), where a 10% imposter rate doubled false alarms (relative to 50% PMT). Though our false alarm rate for 50% PMT was consistent with theirs (around 20%), their false alarm rate for low PMT was much higher (around 45%). Why the different false alarm rates? Papesh and Goldinger's (2014) task differed from ours (and Bindemann et al., 2010) in several respects, including that each face pair involved one large photo and one much smaller photo embedded in a mock ID card. However, an important difference could be that they included time penalties (pauses in the task) for incorrect responses. These penalties (or the absence of them for correct responses) served as feedback throughout the task.

Table 9
Mean Scores on Calibration Statistics for Positive and Negative Decisions in Experiment 2 With (Standard Deviations) and [95% Bootstrap Confidence Intervals]

	"Same person" decisions		"Different people" decisions	
Statistic	3% PMT	50% PMT	3% PMT	50% PMT
C	.052 (.028) [.046, .057]	.034 (.022) [.030, .039]	.369 (.127) [.344, .393]	.039 (.021) [.035, .043]
ANDI	.000	.118	.099	.116
O/U	−.180 (.076) [−.194, −.165]	−.003 (.110) [−.025, .020]	.546 (.184) [.513, .583]	−.033 (.112) [−.055, −.013]

Note. ANDI scores are for data aggregated across participants, due to insufficient incorrect responses for individual-level calculations.

Participants in the 10% PMT condition probably soon learned that imposters were indeed rare and thus lowered their decision criterion, choosing "match" more often and therefore making more false alarms when the rare imposters did appear. This explanation should be tested in future experiments. In applied face matching settings, people do not typically obtain trial-by-trial feedback on their accuracy (indeed, the correct answer is usually unknown), so the false alarm rates in these settings could be more like the low rates that we and Bindemann et al. (2010) observed.

Experiment 2 also confirmed two key findings about the effect of very low PMT on the relationship between confidence and accuracy in face matching. First, overall, confidence is still a good indicator of accuracy, with only slight miscalibration and overconfidence for 3% PMT. Though calibration and resolution were best in the 20% PMT conditions (Experiment 1), they were still quite reasonable for 3% PMT. Second, despite the strong overall confidence–accuracy relationship, a positive–negative asymmetry emerges when the proportion of matches and mismatches is unequal, and this varies systematically with the extent and direction of changes in PMT. Importantly for some applied face matching settings, for very low PMT there is extreme underconfidence for positive judgments, and extreme overconfidence for negative judgments. We discuss this effect further in the next section.

As a final note, one potential limitation to the generalizability of the results from the 3% PMT condition is that the same two mismatching face pairs were presented to all participants. We thought it was important to control the accuracy of the two critical nonmatch trials in this condition, so chose face pairs of midrange difficulty (around 75% accuracy based on Experiment 1). The extent to which our findings extend to applied face matching settings will depend upon the difficulty of the rare nonmatches, which is unknown.

General Discussion

Our goal was to further understand the relationship between confidence and accuracy in unfamiliar face matching, and investigate whether confidence could be used as an indicator of accuracy in applied settings. We examined the overall confidence–accuracy relationship, as well as the relationship for positive and negative decisions. Furthermore, we explored whether these relationships are affected by the task orientation of the decision-maker (look for matches, or look for imposters) and the proportion of mismatching face pairs. We have several key findings.

Confidence–Accuracy Relationship for a 50% Mismatch Rate

If we consider the confidence–accuracy relationships when the proportions of matching and mismatching pairs are equal, confidence appears very promising as an indicator of accuracy. As has been found in studies of eyewitness identification and face recognition (see Brewer & Weber, 2008; Brewer & Wells, 2006; i.e., in the absence of distorting factors), there was a strong overall relationship. However, unlike studies from that field (e.g., Weber & Brewer, 2003, 2004), we also found a strong relationship for both positive and negative decisions. Experiment 1 showed an asymmetry similar to that of Brewer and Wells (2006) in particular, but Experiment 2 suggested that the asymmetry was removed when the matching and mismatching face pairs were equated for difficulty. Therefore, in one-to-one face matching, on average people can be remarkably well-attuned to the chances of being correct, whether the decision is positive or negative. This is especially impressive given our untrained, inexperienced samples.

There are two noteworthy implications from the lack of asymmetry for 50% PMT. First, in eyewitness and face recognition studies, the asymmetry has been attributed to people using different evidence to determine confidence for "old"/positive versus "new"/negative recognition decisions (Lindsay et al., 2013; Weber & Brewer, 2004, 2006). The key idea is that for "old" decisions, confidence can be assessed in a one-to-one mapping between a test face and a specific item in memory, but for "new" decisions, a simple comparison like this cannot be made. While we did not set out to test this hypothesis, our findings accord with it. One-to-one comparisons can be used for both positive and negative decisions in face matching, and we see no asymmetry (when difficulty is controlled).

Second, as an important methodological point, our experiments suggest that studies need to consider the difficulty of targets and lures (e.g., matches and mismatches) when examining calibration. For positive, negative, and overall decisions, the confidence–accuracy relationship may be affected when difficulty is not carefully controlled. Response bias might also be affected, which contributes to overall accuracy.

The Effect of Task Orientation

Experiment 1 examined the impact of task orientation—that is, by what counted as targets and lures for the decision-maker. At the outset, we hypothesized that another possible reason for the positive–negative asymmetry in memory for faces is that people focus on evidence that supports a positive decision, which is usually defined as "old" (i.e., "seen during the study phase"). In face matching this would be most equivalent to "match," but a positive decision could instead be "imposter," depending on task orientation. However, we found that task orientation had little effect on accuracy or its relationship with confidence. This suggests that people perform the task and assess confidence in similar ways for both orientations—even when matches or imposters are more common. We note that a match-orientation might be people's default or standard approach, because people tended to be a little slower under an imposter-orientation, especially for high PMT (see Table 1 and the supplemental materials). Nevertheless, we conclude that in applied face matching settings, a shift in instructions to watch out for imposters would have little impact on the confidence–accuracy relationships. Task orientation may still be important in eyewitness identification and face recognition, where there is a consistent positive-negative asymmetry. This could be addressed in future studies.

The Effect of the Proportion of Mismatching Face Pairs

Both experiments manipulated the proportion of mismatching trials, and found that even when mismatches were more frequent or very rare, overall there was a strong correspondence between confidence and accuracy. However, we found a striking positive-negative asymmetry that varied systematically with the extent and direction of changes in PMT. For high PMT (80%), there was overconfidence for positive decisions and underconfidence plus a flatter calibration curve for negative decisions. As PMT decreased, the confidence–accuracy relationships inverted. Of concern for some applied face matching settings, for very low PMT (3%) there was extreme underconfidence for positive decisions and overconfidence for negative decisions, with calibration curves that were especially flat for positive decisions. These mirrored-shifts in calibration curves align with effects that have been indirectly observed in face memory studies (Brewer & Wells, 2006; Weber & Brewer, 2004 vs. Weber & Brewer, 2006).

PMT affects the asymmetry because the calculations for producing separate calibration curves for positive and negative decisions incorporate both performance rates and the proportion of matching and mismatching trials. This becomes obvious if the scores that form a calibration curve are considered in terms of Bayes's theorem. For example, for positive decisions, a point for a given level of confidence represents the probability that a pair is a match, *M*, given that people identified it as a match, *m*. This probability is given by:

$$p(M|m) = \frac{p(m|M) * p(M)}{p(m|M) * p(M) + p(m|N) * p(N)}$$

where *N* represents nonmatches. Therefore, performance—$p(m|M)$ and $p(m|N)$, or the hit rate and false alarm rate, respectively—can remain the same, but if the proportion of match trials, $p(M)$, greatly changes then the points on a calibration curve will be very different (see Getty, Swets, Pickett, & Gonthier, 1995).

Given the above equation, what is the basis of our observed positive–negative asymmetries for different levels of PMT? Were the different curves at least partially due to changes in performance between conditions (i.e., $p(m|M)$ and $p(m|N)$ for each level of confidence), or did $p(M)$ and $p(N)$ drive the extent of the asymmetries? After all, we did see shifts at least in response bias across PMT conditions. To test this, we used the predicted hit rates and false alarm rates from the best-fitting unequal variance signal detection model for the 20%, 50%, and 80% PMT conditions in Experiment 1 (match-orientation only). These rates are shown in the ROC curves in Figure 7 (left panel), and now include all 12 levels of confidence rather than reducing them to six levels. Each set of rates (i.e., from each condition) was used to generate the calibration curves that would be produced if that same performance occurred under different hypothetical levels of $p(N)$ or PMT: 20%, 50%, and 80%,[5] as shown in the middle and right panels of Figure 7. For example, the set of predicted rates from the 20% PMT condition in Experiment 1 (triangular points in the left panel), was used to generate calibration curves under a hypothetical 20%, 50%, and 80% PMT (light gray to dark gray triangles in the middle and right panels). If the small differences in performance between conditions reflected in the model ROC curves affected the degree of the confidence–accuracy asymmetry that we observed in each condition in Experiment 1, we should see different calibration curves generated from each set of rates. Instead, the three models produce similar calibration curves within each hypothetical PMT level (the same effect occurs if we perform this procedure for Experiment 2). This suggests that the extent of the positive-negative asymmetry was influenced mostly by PMT or $p(N)$, and little by the observed changes in people's performance (e.g., the shifts in response bias across different PMT).

For confidence to be a reliable indicator of accuracy for both positive and negative decisions under low PMT, we would need to drastically change people's performance. Our experiments found that at the aggregate level, as PMT varied, discriminability remained fairly constant but the decision criteria changed, especially under the most extreme 3% PMT. However, these changes did not lead to good confidence–accuracy calibration for positive and negative decisions. Importantly, it is theoretically possible to achieve near-perfect calibration for both decisions under low PMT. We tried optimizing the parameters of the signal detection model to find the performance required to produce perfect calibration under 3% PMT. One solution involves drastically improving discriminability (more than doubling it), producing an ROC curve that closely follows the top-left corner of excellent discrimination. Such high discriminability would probably be difficult for people to achieve in unfamiliar face matching. Another solution is to assume that discriminability is fixed, and adjust the decision criteria. This might seem like a more realistic way to change people's performance, but unfortunately huge shifts in criteria would be required for near-perfect calibration under 3% PMT. The black triangles in Figure 8 (left panel) show the estimated ROC curves from model fitting for the 3% PMT condition, along with the corresponding calibration curves that show an extreme positive–negative asymmetry (middle and right panels). The gray triangles show the shift in criteria required (left panel) to produce near-perfect calibration (middle and right panels). People would need to set much more lenient (lower) criteria, in line with the high proportion of matches. Such a large shift might also be difficult to achieve, and perhaps more importantly, would come at a large cost to the false alarm rate, as can be seen in the ROC curve. In applied settings, this presents the danger that more imposters could be missed. Nevertheless, a compromise in performance may be possible, with criteria set in between the two extremes shown in Figure 8. Thus, calibration for positive and negative decisions might be improved (though not perfect), with a smaller detriment to false alarms. Even better, if discriminability could also be somewhat improved, a reasonable balance might be achieved between the extent of the positive-negative asymmetry and overall false alarms. Given that our participants were untrained and inexperienced with unfamiliar face matching, such a goal seems attainable.

Future face matching research should try manipulating people's response bias and discriminability under low PMT to examine the effects on performance and calibration for positive and negative decisions. Note that under low PMT, our participants set more-stringent (higher) criteria, responding with "mismatch" too often. However, assuming fixed discriminability, for good calibration for both positive and negative decisions, the decision criteria need to change in the opposite direction. Merely informing people of the expected PMT may not be enough. As discussed earlier, Experiment 2 showed comparable hit rates and false alarm rates to that of Bindemann et al. (2010), where unlike in our experiments, partic-

[5] The calibration curves were generated by simply multiplying the predicted hit rates and false alarm rates from the model ROC curve by the appropriate number of targets or lures (respectively) in the hypothetical condition (e.g., 48 matches and 12 mismatches for 20% PMT). This finds the predicted *cumulative number* of responses across the confidence categories for matches and mismatches, which can be converted to the predicted *number* of responses in each confidence category.

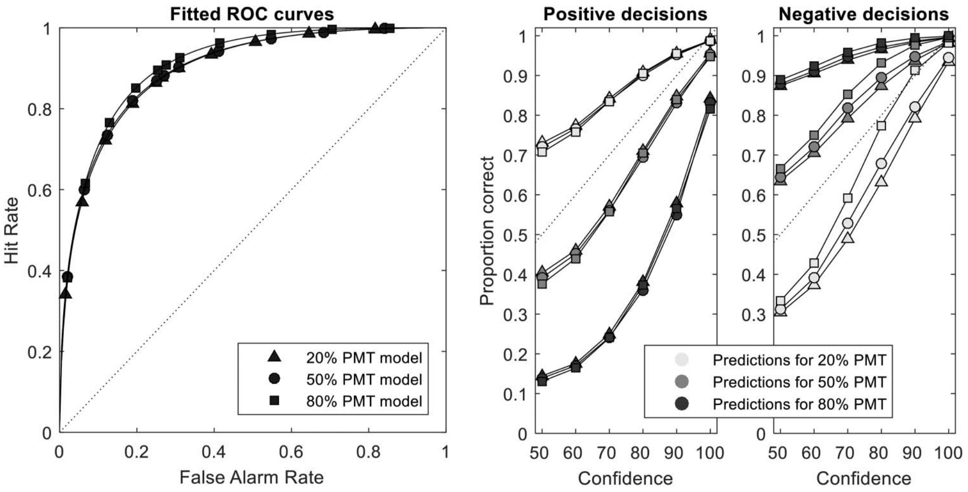

Figure 7. Left panel: Estimated ROC curves from the fitted unequal variance signal detection models of Experiment 1 (match-orientation conditions). Middle and right panels: The resulting confidence–accuracy calibration curves (for positive and negative decisions) from responses generated by the three models when each is applied to a hypothetical PMT of 20%, 50%, or 80%. The models produce similar calibration curves within each hypothetical PMT level.

ipants were informed about the very low (or equal) PMT before the task. For criteria to change, explicit instruction about how decisions and confidence should incorporate PMT may be required. Alternatively, simply training with trial-by-trial feedback on decision-accuracy may reduce the positive-negative asymmetry. Indeed, Papesh and Goldinger (2014) gave participants implicit feedback via the time penalties, and found that false alarm rates doubled for 10% PMT (relative to 50% PMT). This increased false alarm rate is in the correct direction for reducing the positive–negative asymmetry for low PMT, as shown in the ROC curves in Figure 8. Feedback may also help to improve discriminability, particularly for individuals of lower aptitude, as found by White, Kemp, Jenkins, and Burton (2014) for 50% PMT. Research on simple vigilance tasks suggests that sensitivity and response bias can be improved simultaneously via feedback (e.g., Szalma, Miller, Hitchcock, Warm, & Dember, 1999). Importantly, the effect of feedback must be shown to extend beyond such training sessions to performance when feedback is no longer offered, mirroring applied settings. Another avenue for future research is to test the confidence–accuracy relationships for "super recognizers" and specially trained experts who excel at face matching (e.g., Norell et al., 2015; Russell, Duchaine, & Nakayama, 2009).

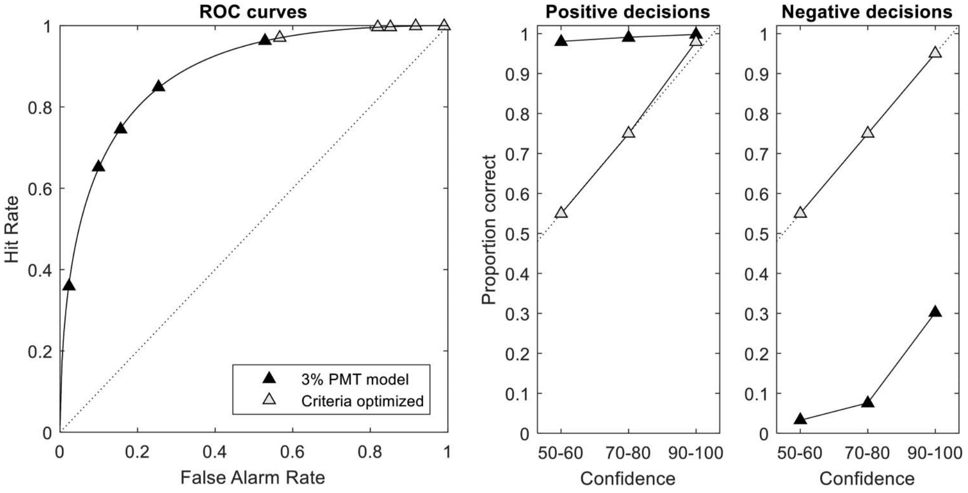

Figure 8. Left panel: Estimated ROC curves from the unequal variance signal detection model of 3% PMT in Experiment 2 (black triangles), and with the decision criteria optimized for near-perfect calibration for positive and negative decisions (gray triangles). Middle and right panels: The resulting confidence–accuracy calibration curves for positive and negative decisions for each set of criteria. Thus, it is theoretically possible to achieve near-perfect calibration by adjusting response bias, but at a cost to the false alarm rate.

For now, our experiments present important baseline findings about the confidence–accuracy relationships for unequal PMT. They suggest that confidence can be useful as an indicator of accuracy—at least in certain respects. Regardless of PMT, confidence tracks accuracy well overall. However, in applied settings with very low PMT, accuracy cannot be so well-predicted from separately considering positive and negative decisions. Performance does not drastically change with shifts in PMT, and so the calibration curves actually mostly reflect the different rates of matches versus mismatches. When imposters are very rare, most positive decisions will indeed be correct, regardless of confidence. However, confidence may offer some useful information for negative decisions, because we found that decisions with confidence of 80 or higher were more likely to be correct than decisions with lower confidence (see Figure 6). To address a question such as: "If a customs officer is 90% confident in a nonmatch decision, how likely is she to be correct?" base rate indeed *should* be taken into account. It is thus useful that calibration curves incorporate PMT information.

More generally, our results also show that researchers need to be aware of the impact of unequal targets and lures when assessing the positive–negative asymmetry. Indeed, our positive and negative calibration curves for 80% PMT in Experiment 1 are similar to those of Weber and Brewer (2006, with 67% "new" trials), with overconfidence for positive decisions and underconfidence plus much flatter calibration curves for negative decisions. When PMT is not 50%, the unequal proportions of targets and lures is reflected by a positive–negative asymmetry (unless there happens to be large shifts in performance in the optimal direction). In such cases, it cannot be assumed that the asymmetry is due to different processes for positive and negative decisions. Eyewitness identification, face recognition, and face matching studies typically involve balanced targets and lures. However, the base rate or proportion of target trials is an important issue for both eyewitness and face matching research, especially where the goal is to investigate when confidence predicts accuracy. Just as matches are more common in many real-world face matching settings, for eyewitness settings, depending on police lineup practices, many lineups may be target-present (for a discussion, see Brewer & Wells, 2006). Thus, our results will be relevant to research in both areas as the impact of base rate is further explored.

The Application of Calibration and Signal Detection Approaches to Face Matching

Our work shows the usefulness of applying calibration and signal detection approaches to face matching for distinct but complementary information. SDT measures and ROC curves offer richer insights into performance than simply examining binary decisions and proportion correct, but cannot directly be used to predict the accuracy of a decision given the reported level of confidence. Instead, this can be easily read off an appropriate calibration curve. Both approaches are needed and will be even more important for examining the effects of feedback and so on, because the confidence–accuracy relationship can remain the same even if discriminability changes, and vice versa (Mickes, 2015; also see Szalma et al., 2006, for a similar argument about assessing performance via both signal detection and diagnostic accuracy measures). As we showed, SDT can also be used to understand the changes in performance required to improve calibration. Unlike calibration curves, ROC curves are independent of PMT or base rate (Wixted & Mickes, 2015), and so are useful for directly comparing performance across different PMT levels.

As a final note to face matching researchers, the standard measures of discriminability and response bias, d' and C should be applied with caution. These measures assume equal variance for the underlying strength distributions, but our asymmetrical empirical ROC curves suggest that this assumption is not necessarily met for one-to-one face matching.

Conclusion

People make errors during unfamiliar face matching decisions. However, confidence offers useful information about their decisions and is promising as an indicator of accuracy. Our work details baseline face matching, and shows that while task orientation may not matter, the proportion of mismatching trials affects performance and the confidence–accuracy relationships. Further research is needed on improving calibration for positive versus negative decisions under low proportions of mismatching trials, which is important for many applied settings. We have demonstrated how signal detection and calibration approaches can be constructively applied together to further understand face matching, and the conditions in which confidence predicts accuracy.

References

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55,* 412–428. http://dx.doi.org/10.3758/BF03205299

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,* 378–386. http://dx.doi.org/10.1037/a0021893

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indices of memory processes not just markers of accuracy. *Applied Cognitive Psychology, 22,* 827–840. http://dx.doi.org/10.1002/acp.1486

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11–30. http://dx.doi.org/10.1037/1076-898X.12.1.11

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

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

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42,* 286–291. http://dx.doi.org/10.3758/BRM.42.1.286

Dunn, J. C. (2010). How to fit models of recognition memory data using maximum likelihood. *International Journal of Psychological Research, 3,* 140–149.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26,* 32–53. http://dx.doi.org/10.1016/0030-5073(80)90045-8

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1,* 19–33. http://dx.doi.org/10.1037/1076-898X.1.1.19

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

Hetter, K., & Cripps, K. (2014, March 11). Who travels with a stolen passport? *CNN*. Retrieved from http://ed.cnn.com/2014/03/10/travel/malaysia-airlines-stolen-passports/

Horry, R., & Brewer, N. (2016). How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145,* 1615–1634. http://dx.doi.org/10.1037/xge0000227

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11,* 211–222. http://dx.doi.org/10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36,* 1–16.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118. http://dx.doi.org/10.1037/0278-7393.6.2.107

Lawrence, M. A. (2013). *ez: Easy analysis and visualization of factorial experiments. R package version 4.2–2*. Retrieved from http://CRAN.R-project.org/package=ez

Lindsay, R. C. L., Kalmet, N., Leung, J., Bertrand, M. I., Sauer, J. D., & Sauerland, M. (2013). Confidence and accuracy of lineup selections and rejections: Postdicting rejection accuracy with confidence. *Journal of Applied Research in Memory & Cognition, 2,* 179–184. http://dx.doi.org/10.1016/j.jarmac.2013.06.002

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.

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

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14,* 364–372. http://dx.doi.org/10.1037/a0013464

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory & Cognition, 4,* 93–102. http://dx.doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376. http://dx.doi.org/10.1037/a0030609

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239–257. http://dx.doi.org/10.1037/a0023007

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences, 60,* 331–340. http://dx.doi.org/10.1111/1556-4029.12660

Olsson, N., & Juslin, P. (2002). Calibration of confidence among eyewitnesses and earwitnesses. In P. Chambres, M. Izaute, & P. J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 203–218). Boston, MA: Kluwer Academic Publishers. http://dx.doi.org/10.1007/978-1-4615-1099-4_14

O'Toole, A. J., An, X., Dunlop, J., Natu, V., & Phillips, P. J. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception, 9,* 16.

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research, 51,* 2145–2155. http://dx.doi.org/10.1016/j.visres.2011.08.009

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

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. *Ergonomics, 40,* 390–399. http://dx.doi.org/10.1080/001401397188224

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 1090–1104. http://dx.doi.org/10.1109/34.879790

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32,* 74–85. http://dx.doi.org/10.1016/j.imavis.2013.12.002

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

Sauerland, M., Sagana, A., & Sporer, S. L. (2012). Assessing nonchoosers' eyewitness identification accuracy from photographic showups by using confidence and response times. *Law and Human Behavior, 36,* 394–403. http://dx.doi.org/10.1037/h0093926

Smith, M., & Ferrell, W. R. (1983). The effect of base rate on calibration of subjective probability for true-false questions: Model and experiment. In P. Humphreys, O. Svenson, & A. Vari (Eds.), *Analysing and aiding decision processes* (pp. 469–488). Amsterdam, the Netherlands: North-Holland. http://dx.doi.org/10.1016/S0166-4115(08)62251-7

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. http://dx.doi.org/10.1037/0033-2909.118.3.315

Szalma, J. L., Hancock, P. A., Warm, J. S., Dember, W. N., & Parsons, K. S. (2006). Training for vigilance: Using predictive power to evaluate feedback effectiveness. *Human Factors, 48,* 682–692. http://dx.doi.org/10.1518/001872006779166343

Szalma, J. L., Miller, L. C., Hitchcock, E. M., Warm, J. S., & Dember, W. N. (1999). Intraclass and interclass transfer of training for vigilance. In M. W. Scerbo & M. Mouloua (Eds.), *Automation technology and human performance: Current research and trends* (pp. 183–187). Mahwah, NJ: Erlbaum.

Van Zandt, T. (2000). ROC curves and confidence judgements in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. http://dx.doi.org/10.1037/0278-7393.26.3.582

Vickers, D. (1979). *Decision Processes in Visual Perception*. London, UK: Academic Press.

Vickers, D. (1985). Antagonistic influences on performance change in detection and discrimination tasks. In G. d'Ydewalle (Ed.), *Cognition, information processing, and motivation* (pp. 79–115). Amsterdam, the Netherlands: North-Holland.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88,* 490–499. http://dx.doi.org/10.1037/0021-9010.88.3.490

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10,* 156–172. http://dx.doi.org/10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy, and response latency. *Applied Cognitive Psychology, 20,* 17–31. http://dx.doi.org/10.1002/acp.1166

Weber, N., Woodard, L., & Williamson, P. (2013). Decision strategies and the confidence–accuracy relationship in face recognition. *Journal of*

Behavioral Decision Making, 26, 152–163. http://dx.doi.org/10.1002/bdm.1750

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE, 10,* e0139827. http://dx.doi.org/10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review, 21,* 100–106. http://dx.doi.org/10.3758/s13423-013-0475-3

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

Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory & Cognition, 4,* 329–334. http://dx.doi.org/10.1016/j.jarmac.2015.08.007

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005, May 26). Cognitive psychology: Rare items often missed in visual searches. *Nature, 435,* 439–440. http://dx.doi.org/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,* 623–638. http://dx.doi.org/10.1037/0096-3445.136.4.623

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110,* 611–617. http://dx.doi.org/10.1037/0033-2909.110.3.611

Received June 23, 2016
Revision received February 8, 2017
Accepted April 3, 2017 ■