

A criterion-placement theory of face matching

Kristen A. Baker^{a,*}, Catherine J. Mondloch^b, Peter J.B. Hancock^c, Markus Bindemann^a

^a School of Psychology, University of Kent, UK

^b Psychology Department, Brock University, Canada

^c Psychology, Faculty of Natural Sciences, University of Stirling, UK

ARTICLE INFO

Keywords:

Face matching
Unfamiliar face identification
Person identification
Individual differences
Decision criterion

ABSTRACT

Face matching is an important applied task that requires binary decisions to pairs of face images to determine whether these depict the same person (an identity match) or different people (a mismatch). While these choices are mutually exclusive, performance for match and mismatch trials appears to be dissociable, which poses a problem for theory development. The current study demonstrates that this dissociation arises from systematic response biases, which reflect individual differences in the placement of decision-making thresholds to distinguish matches from mismatches. When these biases are controlled or partialled out from classification accuracy, reliable associations between match and mismatch identifications are found. This is demonstrated over two experiments with a sample of over 500 participants, several face-matching tests, and a series of data simulations. These findings support a cognitive theory in which individual differences in the placement of decision-making thresholds provide the mechanism by which the identification of face matches and mismatches are linked.

1. Introduction

In our modern world, photographs of faces are ubiquitous. Often, these facial images depict friends and family or famous people. However, through various means we are also exposed to a very large number of images of *unfamiliar* people, who are unknown to an observer. In the personal or social domains, such images can be used to determine whether the same person is depicted in different instances. In security and police settings, the comparison of facial images is used to establish who someone is or to confirm that a person is who they say they are. Because of this importance of facial comparisons in applied settings, *face matching* has been studied extensively in Psychology (e.g., Bruce et al., 2001; Burton et al., 2010; Fysh & Bindemann, 2018). A large body of evidence now shows that this is a difficult and error-prone task not only for laypeople but for professionals in the border control, security, forensic and police sectors (e.g., White et al., 2014; White et al., 2015; Wirth & Carbon, 2017; for a review, see White & Burton, 2022).

Unfamiliar face matching is a challenging task because a person's appearance can change substantially across images or encounters. This within-person variation arises from the complex interplay of a range of

factors, such as temporary changes in facial expression (Bruce et al., 1999; Chen et al., 2011; Mileva et al., 2020), longer-term changes in appearance (Bindemann & Sandford, 2011; Jenkins et al., 2011; Megreya et al., 2013), and variation in viewpoint (Estudillo & Bindemann, 2014; Favelle et al., 2017; Hill & Bruce, 1996), lighting (Hill & Bruce, 1996; Liu et al., 2013), or distortion by camera lens (Noyes & Jenkins, 2017). As a consequence, the same person can sometimes look deceptively different across images, while two different individuals can also look very similar.

In face matching, observers must solve this interplay of within-person variability and between-person similarity to determine whether a pair of face images depicts the same person (i.e., an identity match) or two different people (a non-match or mismatch). These two decisions appear to be linked: If two faces do not depict the same person, then they must depict different people. In turn, one might expect that observers who are adept at identifying when two faces match, are also good at determining when a face pair shows different people. Paradoxically, however, some studies suggest that these task aspects might be unrelated and driven by dissociable cognitive processes. For example, one might expect that face images that are easy to classify as an identity

* Corresponding author.

E-mail address: k.baker-737@kent.ac.uk (K.A. Baker).

<https://doi.org/10.1016/j.cognition.2025.106319>

Received 20 June 2025; Received in revised form 1 September 2025; Accepted 2 September 2025

Available online 27 September 2025

0010-0277/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

match are also distinguished more easily from other faces, but by-item analyses of identification decisions reveal no association in accuracy across both types of face pairings (e.g., Fysh & Bindemann, 2023; Megreya & Burton, 2007; Sauerland et al., 2016).

This finding is remarkable considering strong by-item correlations for matches and mismatches of *familiar* identities, whereby faces that are easier to recognise as an identity match are also easier to reject when they do not match another person (Megreya & Burton, 2007). However, converging evidence for the match-mismatch dissociation in unfamiliar face matching comes from a range of other sources. In factor analysis, for example, individual differences in accuracy on match and mismatch trials load onto separate factors (Bobak et al., 2023), and different developmental trajectories are associated with improvements in performance on match and mismatch trials (Megreya & Bindemann, 2015). Moreover, manipulations that drive improvements in the learning of unfamiliar face identities lead to separate increases in accuracy on match or mismatch trials, but not both (Ritchie & Burton, 2017; for examples see Matthews & Mondloch, 2018; Menon et al., 2015; for converging old/new findings see Baker & Mondloch, 2019; Zhou et al., 2018). Similar dissociations are seen in training effects for unfamiliar face matching, where instructions to attend to some features (e.g., the eyebrow regions) yields improved performance on match trials (Megreya & Bindemann, 2018), while increased attention to other features (e.g., facial marks such as moles and blemishes) improves performance on mismatch trials (Towler et al., 2021).

These effects have been interpreted widely as evidence that face matching is not a unitary process, but that there is more than one route to identification, whereby match and mismatch decisions might reflect separable processes (e.g., Baker & Mondloch, 2019; Bate et al., 2018; Berger et al., 2022; Bindemann & Burton, 2021; Estudillo & Wong, 2022; Fysh & Bindemann, 2023). However, this match-mismatch dissociation is difficult to reconcile with a key observation—both types of decisions rely on a visual comparison of the similarity of faces within the pair. There is some evidence that similarity is used differently to arrive at these decisions, whereby observers might accumulate convergent information from facial features to decide that a face pair is a match, but evaluate a mixture of matching and non-matching information to identify mismatches (Fysh & Bindemann, 2023; see also Bindemann & Burton, 2021). Ultimately, however, accuracy for both matches and mismatches correlates strongly with similarity ratings for face pairings (Fitousi, 2023; Fysh & Bindemann, 2023; Papesh et al., 2018; Rice et al., 2013; Robertson et al., 2017) and so the two tasks must be associated *somehow*. Moreover, observers must approach the identification of both stimulus types in the same manner initially, before an identification decision is reached. The reason for the match-mismatch dissociation is therefore not clear and this poses a great problem for developing a unitary cognitive theory to explain this task. This issue is also important practically as errors on match and mismatch trials are associated with different risks in applied settings. For example, whereas the failure to identify an identity match might prevent a person from legitimately accessing areas or resources, the failure to distinguish non-matching identities would allow for fraudulent access.

The current study sought to investigate the match-mismatch dissociation to provide a theory of face matching that can explain how these two task aspects are linked. One possible explanation comes from the related phenomenon of the *mirror effect* in the recognition memory literature, wherein differences in the accuracy with which individual stimuli (e.g., words) are recognized as ‘old’ (i.e., present in a previous study phase) is mirrored in the accuracy with which previously unseen stimuli are recognized as ‘new’ (i.e., absent in a study phase; e.g., Dobbins & Kroll, 2005; Glanzer & Adams, 1990; Glanzer et al., 1993; Hintzman et al., 1994; Stretch & Wixted, 1998; Wixted, 1992). Although the mirror effect corresponds to individual items in memory and the

match-mismatch dissociation applies to the perceptual matching of faces, the two effects are conceptually related, such that the match-mismatch dissociation is akin to the *absence* of a mirror effect. Notably, the mirror effect is also attenuated by response biases, such as a tendency to categorise items as ‘new’ under conditions of uncertainty. For example, in a recognition memory test for names, a naturally occurring bias to classify names as ‘unfamiliar’ reduces the mirror effect. In turn, the mirror effect re-emerges when this bias is counteracted with a payoff component that differentially rewards responses to familiar and unfamiliar names (e.g., Hilford et al., 2015; see also Glanzer et al., 2009).

There is tentative evidence that the match-mismatch dissociation in face matching might be driven by similar biases to those that attenuate the mirror effect. For example, whereas accuracy on match and mismatch trials is not correlated on a by-item level, weak negative correlations are often observed on a by-subject basis (e.g., Fysh & Bindemann, 2023; Megreya & Burton, 2007). This indicates the presence of response biases among individual observers, whereby some people find the correct balance between match and mismatch decisions difficult to determine and exhibit systematic response tendencies towards one of these outcomes. These biases can be captured by recoding the accuracy of face-matching tasks into hits (correct responses on match trials) and false positives (incorrect responses on mismatch trials) and by converting these variables into the signal detection measures of d' (sensitivity) and criterion (see Stanislaw & Todorov, 1999). Sensitivity can then provide an unbiased measure of an observer’s accuracy across both match and mismatch trials, while criterion reflects the placement of their decision-making threshold to distinguish one type of face pairing from the other (Stanislaw & Todorov, 1999; Summerfield & Egner, 2014).

Some recent approaches have combined such signal detection frameworks with measures of face similarity and decision confidence to explain the match-mismatch dissociation, but this work has focused on item-level analyses that model stimulus characteristics (i.e., the unequal variance signal detection model of face matching; Fitousi, 2023). This work accounts for the match-mismatch dissociation by modelling the classification of stimuli in these two categories as independent signal distributions with their own variances. In turn, the observation that both match and mismatch performance correlates with measures such as the perceived similarity of the two faces in a stimulus pair is taken as evidence that these decisions also have a shared basis, whereby discrete match/mismatch decisions can be transformed onto a continuous similarity signal.

Signal detection analysis can also provide insight into how observer characteristics influence face matching, as *individual differences* between people in criterion placement are linked to accuracy on this task in important ways. For example, individual differences in criterion placement are stable across time and tasks, which indicates that this is a meaningful characteristic of behaviour (Baker et al., 2023). Differences in the sensitivity of individual perceivers also predicts *shifts* in criterion placement. Individuals with lower face-matching ability adopt more extreme criteria when there are more match than mismatch trials, or the cost of mismatch errors is high, and as such show higher levels of liberal and conservative tendencies, respectively (Stabile et al., 2024). Considering the moderating influence of response biases on the mirror effect (e.g., Hilford et al., 2015), this raises the possibility that individual differences in criterion placement might similarly mask a mirror effect in unfamiliar face matching and therefore obscure how the identification of matches and mismatches is linked.

In the current study, this was investigated by capitalizing on the broad individual differences in performance that are observed routinely in unfamiliar face matching (e.g., Baker & Mondloch, 2022, 2023; Baker et al., 2023; Bobak et al., 2023; Fysh & Bindemann, 2018; McCaffery

et al., 2018; Stacchi et al., 2020). First, we demonstrate the match-mismatch dissociation that has been observed in previous studies by showing an absence of a correlation or a negative relationship between match and mismatch accuracy on a by-subject basis. We then examine the extent to which performance on match and mismatch trials is associated when individual differences in criterion placement on this task are controlled or partialled out from match and mismatch accuracy. This was examined first across two experiments with large sample sizes and multiple face-matching tasks. The effect of criterion placement on the relationship between match and mismatch accuracy was then explored further with a series of data simulations.

2. Experiment 1

The purpose of this experiment was to examine whether the classification of identity matches and mismatches is linked when individual differences in face-matching decision-making are taken into account. We first analysed the relationship between accuracy on match and mismatch trials to replicate previous studies that show a positive correlation between these measures does not exist (e.g., Bobak et al., 2023; Fysh & Bindemann, 2023; Kokje et al., 2018; Megreya & Burton, 2007; Sauerland et al., 2016). Consistent with studies of the mirror effect in which associations between measures emerge when response biases are attenuated (Hilford et al., 2015), we then examined whether an association between match and mismatch performance emerges when individual differences in criterion placement for these decisions are controlled for. For this purpose, observers' response criterion was partialled out from face-matching accuracy and performance on match and mismatch trials was then correlated. To establish the generalizability of our findings, this approach was applied to two tasks, comprising the Glasgow Face Matching Test (GFMT; Burton et al., 2010) and the Ambient Image Face Matching Test (AIFMT; Baker et al., 2023).

2.1. Method

2.1.1. Participants

This experiment used data collected by Baker et al. (2023), in which participants completed an online battery of face identification tasks twice, once in each of two testing sessions. The analyses reported here focus on Session 1 data from two tasks in which participants made same/different judgements for unfamiliar face pairs (see below), but Session 2 data was also utilised to provide a measure of test-retest reliability. The analysed sample comprised 249 participants (female: $N = 166$; Age: M

$= 22.56$, $SD = 4.29$). Participants were Brock University students, who received one hour of research credit for their participation, and volunteers from Prolific (www.prolific.com), who were paid a small fee for their participation. As reported in Baker et al. (2023), an additional four participants were tested, but were excluded for failing attention checks.

2.1.2. Stimuli

This experiment employed a shortened version of the GFMT-short originally used by Towler et al. (2019), which comprised 10 identity matches and 10 mismatches (for an example of these trials, see Fig. 1). Both images of a person on match pairings of the GFMT were taken on the same day, with neutral expressions, but with different cameras. In the case of identity matches, two images that were taken approximately 15 min apart with different cameras were combined. For mismatch pairings, images of two different people were paired. These faces were rendered in greyscale, cropped to show only the head, and sized to 350 pixels in width at a resolution of 72 ppi. Image height was adjusted accordingly to preserve the faces' height-to-width ratios.

The AIFMT consisted of 30 match and 30 mismatch trials. The faces in these stimulus pairs were colour images sized to 275×295 pixels at a resolution of 72 ppi, and showed people under ambient viewing conditions (i.e., contained naturally occurring variability) on a grey background. These stimuli were obtained from Brock University's 'Let's Face it' database, the Face and Ocular Challenge Series (Phillips et al., 2011; Phillips & O'Toole, 2014), and using Google and social media (e.g., Instagram and Twitter) searches of identities from the following categories: Celebrities, chefs, or politicians from other countries (e.g., Thea Sofie Loch Naess, Chef Richard McCormick), and minor league athletes. Match trials were created using two images from different days of the same identities. To meet this qualification for google searches, the images must have been from different locations (i.e., noticeably different background) and with different appearance (e.g., hair, lighting, make-up, facial hair, clothing). For social media searches, these differences were verified by upload time stamps. Mismatch trials were created based on physical similarity. Five participants recognized one identity, so these trials were removed from data analysis.

2.1.3. Procedure

Participants completed the experiment online via Testable (www.testable.org). Participants first applied Testable's screen calibration tool to ensure that stimulus size was held constant across the experiment. Next, participants completed a battery of face identification tasks, two of which required same/different responses. Following recommendations

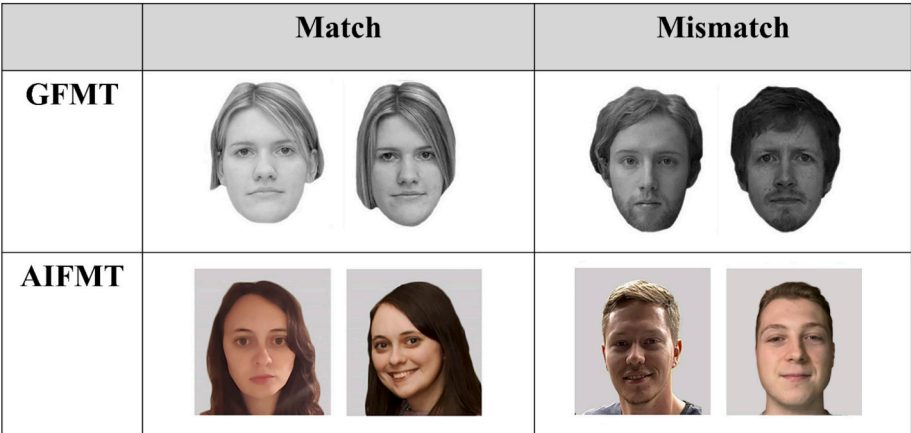


Fig. 1. Note. Fig. 1 provides examples of match and mismatch trials for the Glasgow Face Matching Test (GFMT; top row) and the Ambient Image Face Matching Test (AIFMT; bottom row). For copyright reasons, the mismatch trial provided for the AIFMT is for illustrative purposes and was not used within the task. Models from the images within the images here consented to having their images used within publications.

for individual difference studies (Goodhew & Edwards, 2019), the task order was fixed.¹ On each trial of the GFMT and AIFMT, each face pair remained on screen until a response was registered. Participants were asked to determine whether each face pair depicted the same person or different people, and to indicate their responses by pressing the 'F' key for same-identity pairs and 'J' for different identity pairs. Feedback for responses was not provided.

2.2. Results

All data analyses were performed with SPSS and followed up with Bayesian analyses using JASP 0.18. Observers' face matching accuracy (proportion correct) was first calculated for match and mismatch trials for the GFMT (Match: $M = 0.79$, $SD = 0.20$; Mismatch: $M = 0.77$, $SD = 0.23$) and AIFMT (Match: $M = 0.68$, $SD = 0.17$; Mismatch: $M = 0.66$, $SD = 0.17$). Data was also available from a second test session, at which the GFMT and AIFMT were repeated one week later, and this was used to calculate test-retest reliability using *Pearsons r* correlations. Accuracy on match and on mismatch trials was reliable across the two sessions for the GFMT (match: $r(247) = 0.67$, $p < 0.001$; mismatch: $r(247) = 0.58$, $p < 0.001$) and the AIFMT (match: $r(247) = 0.69$, $p < 0.001$; mismatch: $r(247) = 0.69$, $p < 0.001$).

We then calculated the correlation between match and mismatch accuracy from Session 1 for each test (see Fig. 2). Match and mismatch accuracy were unrelated on the GFMT, $r(247) = -0.05$, $p = 0.44$, $BF_{10} = 0.11$, and correlated negatively on the AIFMT, $r(247) = -0.29$, $p < 0.001$.

0.001, $BF_{10} > 100$. Thus, observers who were good at making match decisions on these two tests do not appear to be the same observers who achieved high mismatch accuracy, consistent with an accuracy dissociation for these measures.

We then examined the relationship between performance on match and mismatch trials when individual differences in decision-making thresholds are partialled out. For this purpose, the accuracy data was converted into signal detection measures of sensitivity, or d' , and criterion, or c (for the calculation of these measures see Stanislaw & Todorov, 1999), using hits (correct responses on match trials) and false positives (incorrect responses on mismatch trials). This revealed GFMT d' values of $M = 1.73$ ($SD = 0.91$) and criterion values of $M = -0.04$ ($SD = 0.48$), and AIFMT d' values of $M = 0.97$ ($SD = 0.66$) and criterion values of $M = -0.02$ ($SD = 0.45$), respectively. These values were previously reported to have good test-retest reliability (i.e., criterion: both $rs > 0.52$, $ps < 0.001$; d' : both $rs > 0.53$, $ps < 0.001$; Baker et al., 2023).

We calculated residual scores to examine the relationships between match and mismatch accuracy after controlling for bias (criterion). To do so, we created averages of each participant's criterion score from the GFMT and the AIFMT using equal weightings (i.e., $\frac{GFMT\ criterion + AIFMT\ criterion}{2}$). This makes good sense as the relationship between criterion on the GFMT and AIFMT is robust, $r = 0.59$, $p < 0.001$ (Baker et al., 2023). Applying the average of these, rather than task-specific scores, reduces the risk of controlling for additional variance that is related to the task and not specifically to criterion. Separately for each task, we then created a residual score for the match trials and

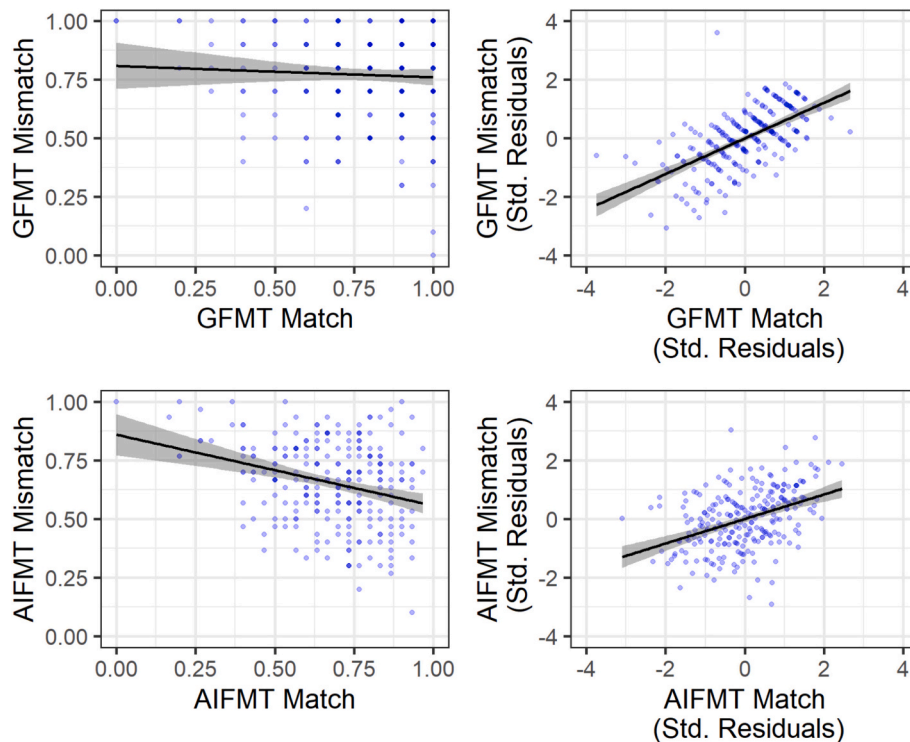


Fig. 2. Note. Fig. 2 depicts the relationships between accuracy on match and mismatch trials for the GFMT (top row) and the AIFMT (bottom). The graphs on the left depict the unadjusted percentage accuracy data. The graphs on the right show a positive association of match and mismatch performance after controlling for criterion.

¹ Task order is recommended to be fixed, not counter-balanced, in individual difference studies. As task performance can differ because of task order, counterbalanced scores would be confounded with this additional variability (Dale & Arnell, 2015). This confound is easily removed by using a fixed task order.

mismatch trials, controlling for criterion which was averaged across both tasks. This was completed by setting either the match or mismatch trials as the outcome variable and average criterion as the predictor. The standardized residual scores were then saved as a new variable. These scores also exhibited good test-retest reliability for GFMT match trials, $r(247) = 0.43$, $p < 0.001$, and mismatch trials, $r(247) = 0.47$, $p < 0.001$, and for AIFMT match, $r(247) = 0.59$, $p < 0.001$, and mismatch trials, $r(247) = 0.41$, $p < 0.001$.

The data of most interest were the match-mismatch correlations of these residualised scores. As can be seen in Fig. 2, after controlling for participant bias (criterion) in this way, positive correlations were observed between match and mismatch accuracy on the GFMT, $r(247) = 0.61$, $p < 0.001$, $BF_{10} > 100$, and the AIFMT, $r(247) = 0.42$, $p < 0.001$, $BF_{10} > 100$.² This data is available on OSF (<https://osf.io/tmvhp>).

2.3. Discussion

The match-mismatch dissociation in face matching poses a profound conundrum for theory development, as it is unclear how performance for identity matches and mismatches can be reconciled as observers attempt to resolve both trial types within the same task. This experiment replicates the dissociation in match-mismatch performance on the GFMT. It also demonstrates a negative correlation between these trial types on the AIFMT. Both findings are consistent with other studies in this domain, which have either shown no or negative correlations between these measures (e.g., Bobak et al., 2023; Fysh & Bindemann, 2023; Kokje et al., 2018; Megreya & Burton, 2007; Sauerland et al., 2016). One interpretation of these results is that distinct abilities are required to solve identity match and mismatch face pairings, where different individuals are good at resolving these seemingly separable aspects of this task. However, the negative correlations that are often observed between match and mismatch accuracy also hint at the presence of systematic response biases in this task. Experiment 1 shows that when such biases are captured by converting performance into criterion, these can be removed from the accuracy data. When individual differences in decision-making during face matching are controlled for in this way, moderate to strong positive correlations emerge between match and mismatch performance.

These findings are consistent with observations of the mirror effect in recognition memory tasks, which also show associations between 'old' and 'new' responses to previously seen and unseen stimuli once participants' response biases are addressed (e.g., Hilford et al., 2015). In the face-matching domain, this is an important finding as it indicates that match and mismatch decisions are based on shared underlying processes. This indicates that a single theory can explain face identification on match and mismatch trials and highlights criterion placement as the important moderating factor. It also provides parsimony between the match-mismatch dissociation and the mirror effect in recognition memory.

3. Experiment 2

Experiment 1 demonstrates that the match-mismatch dissociation arises from individual differences in the placement of an observer's decision-making thresholds to distinguish one type of face pairing from the other. When the decision-making biases that are introduced by this

variation are removed, accuracy for match and mismatch identification correlates. In this experiment, we sought to replicate these effects with a different data set, reflecting face matching performance on the GFMT and the Kent Face Matching Test (KFMT; see Fysh & Bindemann, 2018). Similar to the GFMT, the KFMT has been used extensively in this domain and shows dissociable performance on match and mismatch trials on a by-item basis, while small negative correlations are observed on a by-subject basis (Fysh & Bindemann, 2023). The KFMT therefore provides a suitable test for further examining the theory that individual differences in criterion placement cause the match-mismatch dissociation, while inclusion of the GFMT serves as a reference point for comparison with Experiment 1.

3.1. Method

3.1.1. Participants

The participants comprised 299 white undergraduate students (female: $n = 245$; Age: $M = 20.25$, $SD = 4.98$) from the University of Kent. These participants took part in a larger study containing a battery of tests, which included the GFMT and KFMT that are analysed here, and were paid for their time. Four additional participants were excluded for recognising identities used in one of the other tests ($n = 1$) and inattention (i.e., repeatedly pressing the "S" key through an entire task: $n = 2$; extremely short RTs: $n = 1$).

3.1.2. Stimuli

The short version of the GFMT was employed for this experiment (see Burton et al., 2010), which comprised 20 identity matches and 20 mismatches. As shown in Fig. 3, the size, colour and format of the images, as well as the way in which trials were created were consistent with that reported in Experiment 1. The short version of the KFMT was also used, consisting of 20 matches and mismatches (see Fysh & Bindemann, 2018). The KFMT face pairs consist of an image taken from student identity cards and a passport-style face portrait. The student-ID photos were taken at least three months prior to the face portraits and were not constrained in pose, expression or camera device. The passport-style face portraits were all taken with the same camera, in a frontal pose and with a neutral expression. The student-ID photos were shown at a size of 142×192 pixels, while face portraits were sized to 283×332 pixels at a resolution of 72 ppi.

3.1.3. Procedure

This experiment used data collected for a larger study. Participants completed a battery of tasks, which also included the Cambridge Face Perception Test (CFPT; Duchaine et al., 2007), Cambridge Car Perception Test (CCPT; Yang et al., 2017), Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), Models Memory task (MMT; Bate et al., 2018), and a new face matching task. These tasks were presented in the following order: All participants first completed the new face matching test. For the remaining tasks, there were two different orders, which were randomly allocated to participants. Half of the participants completed the tasks in the order of GFMT, CFMT+, CFPT, MMT, KFMT, whereas the remaining participants completed these tasks in the reverse

² If we use criteria from one test to compute residuals on another, the resultant correlations are naturally reduced because of the additional measurement noise. For example, when task-specific criterion scores from Session 2 are used for regression on Session 1 performance, the match/mismatch relationship is significant for the GFMT, $\rho(208) = 0.18$, $p = 0.008$, and approaches significance for AIFMT, $\rho(208) = 0.12$, $p = 0.09$. This is likely an underestimate of the relationship as different trials were used in Session 1 and Session 2 for each test (see Baker et al., 2023).

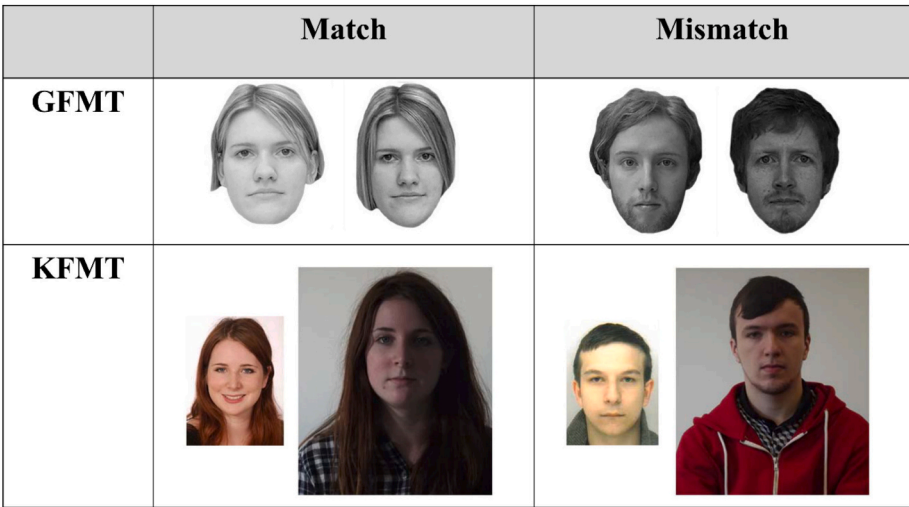


Fig. 3. Note. Fig. 3 provides examples of match and mismatch trials for the Glasgow Face Matching Test (GFMT; top row) and the Kent Face Matching Test (KFMT; bottom row). Models from the images within the images here consented to having their images used within publications.

order. On each trial of the GFMT and KFMT, a fixation cross was shown for 1 s and a pair of faces was then shown until a response was registered. Participants were asked to determine whether the face pairs depicted the same person or different people, and to indicate their response by pressing the S key for ‘same’ or D key for ‘different’. Feedback for accuracy was not provided to participants.

3.2. Results

As in Experiment 1, all data analyses were performed with SPSS and followed up with Bayesian analyses using JASP 0.18. Observers’ face matching accuracy was first calculated for match and mismatch trials for the GFMT (Match: $M = 0.85$, $SD = 0.14$; Mismatch: $M = 0.65$, $SD = 0.20$) and KFMT (Match: $M = 0.69$, $SD = 0.16$; Mismatch: $M = 0.60$, $SD = 0.18$). Split-half reliability estimates for these data show moderate correlations for match trials, $r(297) = 0.55$, $p < 0.001$, and mismatch trials

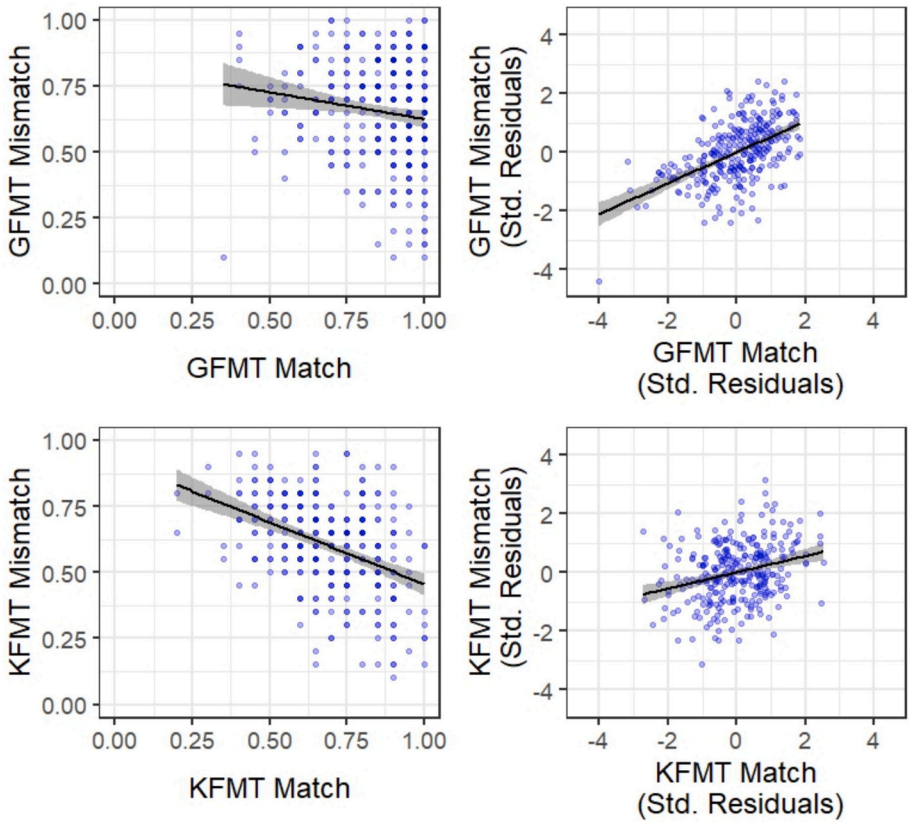


Fig. 4. Note. Fig. 4 depicts the relationships between accuracy on match and mismatch trials for the GFMT (top row) and the KFMT (bottom). The graphs on the left depict the unadjusted percentage accuracy data. The graphs on the right show a positive association of match and mismatch performance after controlling for criterion.

of the GFMT, $r(297) = 0.62, p < 0.001$. This was also observed for match, $r(297) = 0.52, p < 0.001$, and mismatch trials of the KFMT, $r(297) = 0.54, p < 0.001$.

Performance on match and mismatch trials was then correlated using Pearson r for each test. As shown in Fig. 4, match and mismatch accuracy correlated negatively on the GFMT, $r(297) = -0.15, p < 0.001, BF_{10} = 1.85$, and on the KFMT, $r(297) = -0.42, p < 0.001, B_{10} > 100$. Thus, observers who were good at making match decisions on these two tests do not appear to be the same observers who achieved high mismatch accuracy. This is consistent with Experiment 1 and is indicative of differences in criterion placement across participants.

As in Experiment 1, we then examined the relationship between performance on match and mismatch trials when individual differences in decision-making thresholds are partialled out. For this purpose, the accuracy data was converted into signal detection measures of d' and c (for the calculation of these measures see Stanislaw & Todorov, 1999). This revealed GFMT d' values of $M = 1.65 (SD = 0.80)$ and criterion of $M = -0.36 (SD = 0.47)$, and KFMT values of $d' M = 0.85 (SD = 0.56)$ and criterion $M = -0.14 (SD = 0.43)$, respectively. These values also exhibited good split-half reliability for criterion (both $r_s > 0.62, p_s < 0.001$) and adequate split half reliability for d' (both $r_s > 0.29, p_s < 0.001$).

We then calculated residual scores to examine the relationships between match and mismatch accuracy after controlling for bias (criterion). Considering the strong relationship between criterion on the GFMT and the KFMT, $r(297) = 0.55, p < 0.001, BF_{10} > 100$, grand averages were calculated of each participant's criterion score from both tests using equal weightings. These were then used to create a residual score for both match and mismatch trials that controls for average criterion. Once again, these residual scores exhibited high split-half reliability for GFMT match trials, $r(297) = 0.47, p < 0.001$, and mismatch trials, $r(297) = 0.46, p < 0.001$, and for KFMT match, $r(297) = 0.31, p < 0.001$, and mismatch trials, $r(297) = 0.30, p < 0.001$.

The data of most interest were the match-mismatch correlations of these residualised scores. As can be seen in Fig. 4, after controlling for criterion in this way, positive correlations were observed between match and mismatch accuracy on the GFMT, $r(297) = 0.53, p < 0.001, BF_{10} > 100$, and the KFMT, $r(297) = 0.28, p < 0.001, BF_{10} > 100$. This data is available on OSF (https://osf.io/cj428/?view_only=39ee06e2758e45bfa7f60e5712526fb1).

3.3. Discussion

The results provide a direct replication of Experiment 1. Negative correlations were observed between match and mismatch accuracy on the GFMT and the KFMT. This indicates the presence of systematic response biases, whereby observers are biased towards match or mismatch decisions in their attempts to solve these tasks. Again, this interpretation is strengthened by a correlation of criterion values across the two tests, which suggests that these biases are stable at the level of the individual and show transfer across different face-matching tests (Baker et al., 2023). Crucially, when these biases are removed from the accuracy data, reliable positive associations between performance on match and mismatch trials are found for the GFMT and KFMT. This indicates that the relationship between the identification of match and mismatch trials is masked by individual differences in the placement of

decision-making thresholds.

4. Simulation 1

Thus far, we have established that match and mismatch accuracy correlate when individual differences in criterion placement are partialled out from response accuracy. This paves the way for a cognitive theory that can link identification of match and mismatch trials to explain how face matching is solved. However, without a direct manipulation of criterion, the observed correlation of match and mismatch accuracy could be explained by other factors besides individual differences. In the following simulations, we will address possible alternative explanations by attempting to reproduce our findings. The first simulation aims to provide a direct replication of the results from Experiment 1 and 2.

4.1. Method

Here, and in each subsequent simulation, we use the terms ability and bias to distinguish these from the signal detection measures of sensitivity and criterion, which have different units and, in the case of criterion, also have a different sign. For example, while a participant exhibiting a match bias would have a negative criterion value in Experiments 1 and 2, a negative bias will make participants worse at match trials and better at mismatch trials in the simulation. Furthermore, in the simulations, ability is a property of the participant alone, whereas a participant's sensitivity in experiments would be dependent on their ability and the difficulty of the task.

For the first simulation, we generated data for a deterministic model by drawing from separate truncated normal distributions (Luong, 2025)³ to act as participants' ability and bias, and also for stimulus item difficulty. Generating data for these parameters separately means that a simulated individual with high or low ability can have any bias—logic that is consistent with Baker et al. (2023) and with assumptions in signal detection theory.

Then, match and mismatch trials were simulated with a normal distribution of difficulty. A match trial is scored correct if a simulated participant's combined ability and bias (i.e., ability + bias) exceeds the trial difficulty, and a mismatch trial is correct if their combined ability and bias (i.e., ability - bias) value exceeds the trial difficulty. Softening this decision threshold, for example by using a sigmoid probability function, adds complexity and random variance but does not alter the form of the results. We first applied these values to replicate the results of Experiment 1 and 2. We looked at accuracy when bias was allowed to vary, followed by performance after controlling for bias.

4.2. Results

The results of these simulations are shown in Fig. 5. When bias was allowed to vary, as is the case with the raw data in Experiments 1 and 2, performance on match and mismatch trials echoes the negative relationship between match and mismatch accuracy shown in Experiment 2, $r = -0.06, p = 0.59$. We then looked at the relationship between match and mismatch scores after holding individual differences in bias constant. As shown in Fig. 5, this also replicated the findings from Experiment 1 and 2, wherein a positive relationship between match and

³ The truncated Gaussian distributions ensure that the variables for participant ability and item difficulty variables cannot go past 1 or 0. For example, if ability has a $M = 0.65$ and $SD = 0.10$ (capped at 0.20), the data would range between 0.45 and 0.85, and would peak approximately at 0.65.

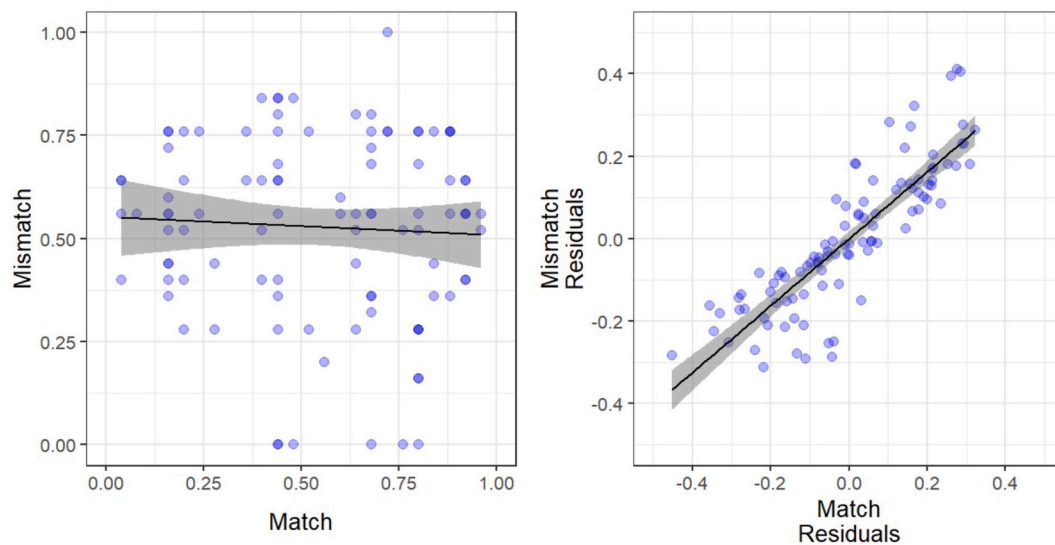


Fig. 5. Note. Fig. 5 provides a Scatter plot of simulated participants' match and mismatch accuracy when bias freely varies (Left) and when bias is controlled (right).

mismatch accuracy emerges after accounting for bias, $r = 0.85$, $p < 0.001$. The correlations are high because, once individual variations in bias are accounted for, the only remaining source of variance is the random normal distribution of item difficulties. Increasing the number of items reduces the average effect of this variation, which increases the correlation further. Adding other sources of variance, such as some uncertainty around each match decision, reduces the correlation.

4.3. Discussion

This simulation replicates the findings with human participants in Experiments 1 and 2, by demonstrating that match and mismatch performance can be correlated negatively when bias values vary freely, but correlated positively when bias is controlled. This was observed under conditions in which match and mismatch trials were simulated with a normal distribution of difficulty, while the performance of individual participants was derived from two normal distributions reflecting ability and bias. We also suggest that item difficulty and the number of trials might be moderating factors, whereby increases in the number of items stabilizes the relative influence of difficulty values for each item, and therefore strengthens the correlation between match and mismatch performance when bias is controlled.

5. Simulation 2

One approach to seeking understanding of the processes involved in face matching and recognition has been to ask participants to do a variety of tasks and then perform a principal components analysis (PCA) on the set of results. This typically produces two components, one of which loads onto matches, or hits in a memory task, with the other loading onto mismatches or correct rejections (e.g., Bobak et al., 2019a, 2019b, Hancock et al., 1996a, 1996b). This has been taken as evidence for there being two different processes underpinning both of matching and recognition memory of unfamiliar faces. Given the results of Simulation 1, this second simulation seeks to explore how variations in individual bias and ability affect the outcome of PCA. We first simulated face matching under conditions of greater variance in participants' bias than ability. This models a scenario that accentuates the impact of

individual differences in criterion on human face-matching performance and should lead to a strong influence of response bias and a negative correlation of match and mismatch scores. We then modelled how this outcome changes when there is equivalent variability in ability and bias, and more variability in ability than bias.

5.1. Method

In each simulation, four different data sets were generated using the same approach as in Simulation 1, akin to behavioural data from four different unfamiliar face-matching tasks. Stimulus items and their associated difficulty values varied across these four simulated data sets, but the simulated participant's ability and bias remained constant. As in the previous simulation, average item difficulty and participant ability were set to a $M = 0.65$ for each of these data sets and simulated participants got a match trial correct if the sum of their ability and bias exceeded the item's difficulty. Values were first simulated for when there is more variance in participants' bias than ability, then with equivalent variability in ability and bias, and finally with greater variability in ability than bias.

5.2. Results

Fig. 6 shows the results of the three different simulations. The plots each result from one of the four sets of match and mismatch scores from each simulation. The PCA results are derived from all four sets of scores for each simulated participant. To simplify the presentation, we report the average values for the loadings of match and mismatch trials, without and with varimax rotation.

As shown in Fig. 6A, when variability in simulated values are greater for bias ($SD = 0.10$) than ability ($SD = 0.05$), match and mismatch scores correlate negatively, $r = -0.67$, $p < 0.001$. In the unrotated PCA, the first component, which captures the larger source of variance, corresponds to bias. This is reflected by the positive component loadings for match scores and negative component loadings for mismatch scores. The second component corresponds to ability. This is reflected by the positive component loadings for both match and mismatch scores. However, when varimax rotation is applied to these data to aid interpretability,

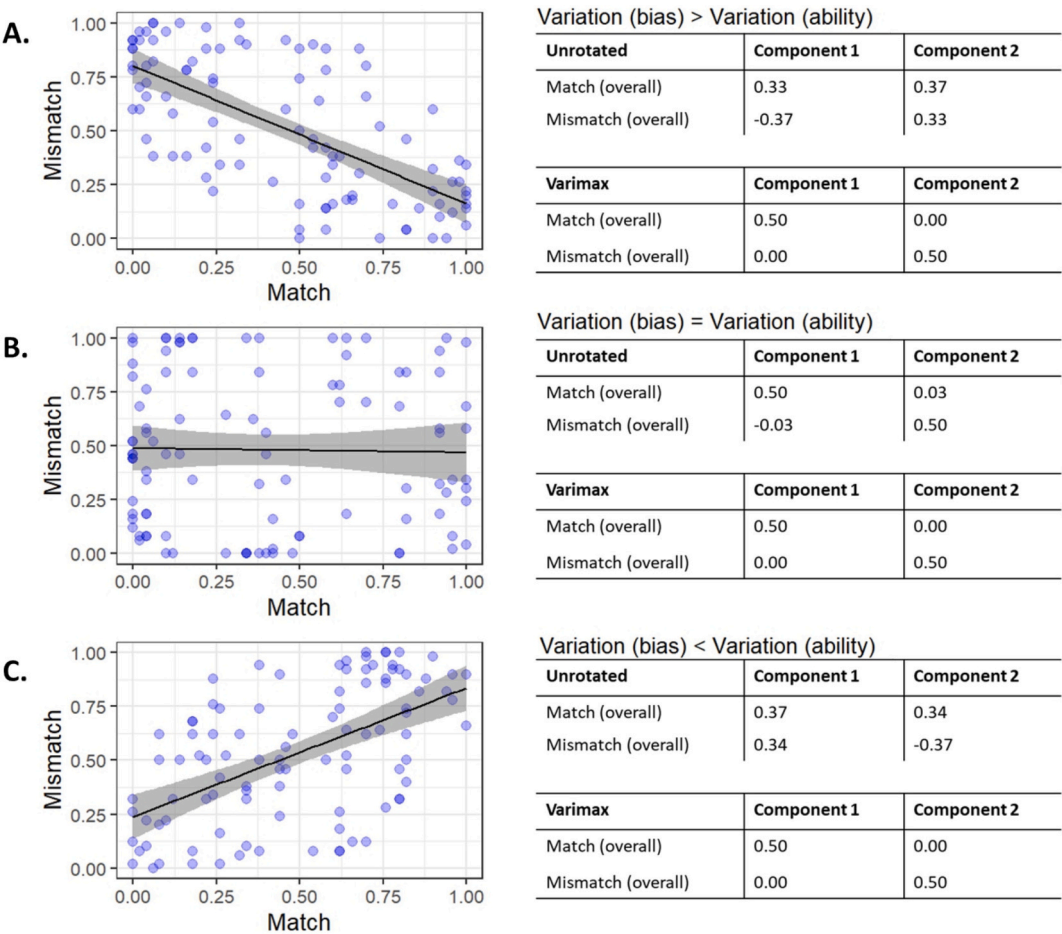


Fig. 6. Note. Fig. 6 provides plots of simulated mismatch vs match accuracy and PCA results for different ratios of variation in ability and bias. The values reported for match and mismatch performance in each PCA reflect an average value. This is because all values in each match and mismatch measure were consistent. Indeed, the maximum amount that any individual match or mismatch score deviated from the reported value above was less than ± 0.10 .

matches and mismatches separate, with one component loading onto accuracy on match trials and the other loading onto mismatches. Varimax rotation aids interpretability by seeking to align the orthogonal PCA axes with input variables and succeeds in aligning one with match and one with mismatch scores.

As shown in Fig. 6B, when the simulation has equal variation in bias and ability ($SDs = 0.10$), there is little correlation between match and mismatch scores, $r = -0.02$, $p = 0.84$. The unrotated PCA solution is somewhat unstable, but here it has settled on one component for match scores (component 1) and the other for mismatch scores (component 2). The varimax rotated solution remains as before, separating match and mismatch components. Finally, when the simulation has less variation in bias ($SD = 0.05$) than that for ability ($SD = 0.10$), as is the case for Fig. 6C, there is a strong positive correlation between match and mismatch performance, $r = 0.55$, $p < 0.001$. The unrotated PCA detects that ability carries the most variance in scores, as is reflected by component 1 which has positive component loadings on both match and mismatch scores. The second component captures variations in bias. This is reflected by positive component loadings for match scores and negative loadings for mismatch scores. Again, the varimax rotated solution separates match and mismatch components.

5.3. Discussion

The scatterplot results make it clear that it is possible to produce positive, zero or negative correlations between match and mismatch scores by adjusting the relative amount of variability in bias and ability. When ability dominates, scores are positively correlated; when bias dominates, the correlation is negative. While the size and direction of any observed correlation does not say anything about the number of cognitive processes involved in unfamiliar face matching, it provides comprehensive evidence of how the spread of individual differences in bias and ability might influence the nature of this correlation. Essentially, observing that an individual is better at match trials than mismatch trials is entirely equivalent to saying they have a bias towards saying match. This might be because there are two or more different underlying cognitive processes and an individual is stronger at identifying match/mismatch trials, or because they are biased towards saying match/mismatch (e.g., because their decision thresholds are placed in such a way that less evidence is needed to say match), or some combination of the two.

The unrotated PCA results behave exactly as they should, identifying the two underlying independent sources of variance in the data and allocating the first component to the source with greater variance. When the two sources of variance are equal, the unrotated PCA can settle on separate components for match and mismatch scores. Applying varimax rotation, which is commonly done to aid interpretability, always generates separate components for match and mismatch scores. These results suggest that the patterns that appear in human data will depend on the range of abilities and the range in biases in the tested population. Thus, if the sample includes individuals who are especially good or bad at the task, then a positive correlation between match and mismatch scores is likely. Similarly, the effect of variable criterion will also depend on how accurate the participants are. When sensitivity is high, as is represented by a large separation of signal and noise distributions, criterion placement should not matter and there should be a positive relationship between match and mismatch performance. When sensitivity is poor, as is represented by partially overlapping signal and noise distributions, criterion placement should matter (e.g., a placement that yields more hits will also yield more false alarms) and this might lead to a negative relationship between match and mismatch performance.

Overall, these results extend the findings from our experiments and simulations, by demonstrating how individual differences in sensitivity and criterion – or ability and bias – interact and suggest that both are important for understanding face matching.

6. Simulation 3

While the experiments and simulations above have assessed individual differences in performance at the observer level, it is very possible that individual differences at the item level might also play a role in the dissociation between hits and correct rejections. For example, while observers are likely to determine that two images of the same distinctive face are the same identity in a match trial, a bias akin to those explored above might be elicited if an experimenter could find a suitable, similar-looking identity to pair with that face in a mismatch trial. However, such biases at the item level cannot exist in matching studies, as a given face pair either is or is not a match. The stimulus pair itself cannot be both. Thus, biases at the stimulus level of matching tasks, might merely reflect that some face pairs are just easier than others, reflecting the similarity of the faces depicted (e.g., [Fysh & Bindemann, 2023](#); for similar effects shown by somewhat increasing the distinctiveness of faces see [McIntyre et al., 2013](#)). As such, to even consider the relative contribution of variability in criterion at the item level on face identification choices in humans, one must use an old/new task.

There are two critical reasons why it is feasible that trial-level biases might elicit similar effects to those shown in individual observers: First, item-level biases are readily found in old/new recognition tasks. For example, the tendency for such a bias to occur at the item level, such as when an item is more likely to be declared as old regardless of whether it had or had not been seen before, was termed “context free familiarity” by [Vokey and Read \(1992\)](#). Second, it has long been observed that old-new memory tasks that investigate face identification elicit a dissociation similar to that which is found in face matching tasks (e.g., [Hancock et al., 1996a, 1996b](#); [Megreya & Burton, 2007](#); [Vokey & Read, 1992](#)). In these studies, items that are correctly remembered when present in one experiment (i.e., are “old”) are not necessarily correctly rejected when acting as foils (lures) in another experiment (i.e., are “new”). Collectively, this suggests that such item-level biases do exist in unfamiliar face identification and might influence the dissociation found in old/new tasks. This final simulation explores how variation in bias across items towards being declared old or new interacts with variations in participants’ bias towards the decision. We hypothesized that both item-level and participant-level biases would influence the relationship between performance on old and new trials in the same way that participant-level biases and criterion influenced the relationship between match and mismatch performance.

6.1. Method

As with the previous simulations, participants were modelled as having a range of abilities and biases, while items were also modelled as having a range of difficulties and biases. In doing so, accuracy is not only dependent on participants’ ability and the item difficulty, but also on each participant’s bias and each item’s bias. Participants’ ability was simulated to be normally distributed and centred on 0.65. Participants each had a bias, also drawn from a normal distribution centred on zero, wherein positive values reflect a greater tendency to classify stimuli as old. Item difficulty was simulated to be normally distributed and centred on 0.65. Items each had a bias, also drawn from a normal distribution centred on 0, where a positive bias makes it more likely to be identified as old. Simulated participants were correct on old trials if their

combined ability and bias (i.e., participant ability + participant bias) exceeds difficulty (i.e., item difficulty - item bias). Simulated participants were correct on new trials if their ability and bias (i.e., participant ability - participant bias) exceeds difficulty (i.e., item difficulty + item bias). Since exactly the same items appeared in both old and new trials, we prevented the model from being completely deterministic by adding a small amount of noise (Gaussian distribution, $M = 0$, $SD = 0.05$) to each decision (i.e., $[\text{ability} \pm \text{bias}] + \text{noise}$). We conducted four variations of this simulation to create the following models: (I) A model without bias at the participant or item level, (II) with bias at the participant level ($SD = 0.05$) and without bias at the item level, (III) without bias at the participant level and with bias at the item level ($SD = 0.05$), and (IV) with bias at both participant ($SD = 0.05$) and item levels ($SD = 0.05$).

6.2. Results

The output from the four simulations can be seen in Fig. 7, with analysis by items on the left and by participants on the right. When there is no variability in bias at either the participant or item level (A and B),

there are strong positive correlations between hits and correct rejections ($r_s > 0.91$, $p_s < 0.001$). Thus, at the participant level, individuals who do well at remembering which items were present at study are also good at rejecting those that were not. At the item level, an item that is well-remembered when present is well-rejected when not. When bias varies only at the participant level (C and D), there is no association between hits and correct rejections at the participant level ($r = -0.15$, $p = 0.13$) but a strong positive correlation at the item level ($r = 0.92$, $p < 0.001$). When bias only varies at the item level (E and F), while there is a strong positive relationship between hits and correct rejections at the subject level ($r = 0.87$, $p < 0.001$), but the correlation is non-existent at the item level ($r = -0.11$, $p = 0.60$). Finally, when variability in bias is present at both the participant and the item levels (G and H), performance is not correlated either for items ($r = -0.08$, $p = 0.71$) or participants ($r = 0.21$, $p = 0.15$). The strength of all correlations depends on the balance between variability in criterion and variability in ability/difficulty. The simulation code is available on osf (<https://osf.io/5wk2z/>).

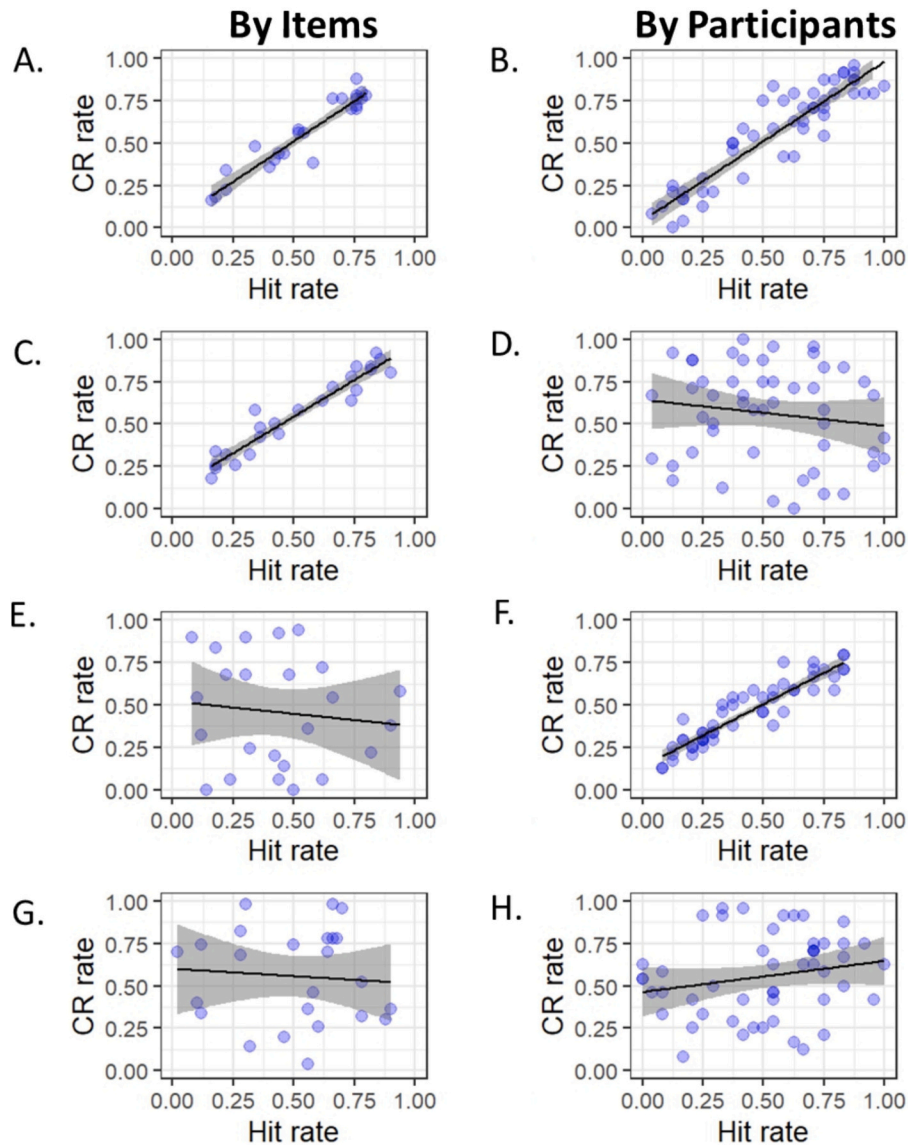


Fig. 7. Note. Fig. 7 provides plots for the relationships between old (hits) and new (correct rejections; CR) responses with and without variability in criterion at the item (left) and participant (right) levels. A-B had no variability in bias at either level. C-D had variability in bias only at the participant level. E-F had variability in bias only at the item level. G-H had variability in bias at both levels.

6.3. Discussion

In this simulation, the relationship of hit and correct rejection rates were modelled by contrasting variability in criterion at the participant and item level. This is an important simulation as it shows that just as individual differences in observers' criterion can drive the dissociation of match and mismatch performance, it also can drive the dissociation of old/new performance. Indeed, this simulation also shows that item-level biases can elicit a similar dissociation. Together, differences in both item- and individual-level biases can account for the long-observed lack of correlation between hit and correct rejection rates in old/new memory tasks. This is critical as, while item-level biases cannot feasibly exist in matching tasks, biases in both items and in criterion can exist in many face identification tasks, and in many applied settings. Notably, the correlations at the item level and the participant level are each dependent on the amount of variation in the bias at the corresponding level, and as such it is possible to get a correlation at the item- but not the participant-level, or vice versa. There are of course other sources of variance that will affect the observed item-level correlations. For example, participants tend to remember unfamiliar faces that resemble those that they do already know (Hancock, 2021; for the influence of similarity in siblings see Strathie et al., 2022). Regardless, the findings here suggest that, much like the match-mismatch dissociation shown in the previous studies and simulations, the dissociation in hits and correct rejections that is often observed in old/new recognition tasks is also influenced by bias—an effect that was hinted at within the recognition memory literature (Hilford et al., 2015).

7. General discussion

Face matching requires binary decisions to pairs of images, to determine whether the images depict an identity match or a mismatch. These two choices are mutually exclusive so one might expect that observers, who are adept at identifying when a matching face pair displays images of the same identity, are also good at determining when a mismatching face pair shows different people. Contrary to this reasoning, performance for match and mismatch trials is typically unrelated or correlated negatively, whereby observers who are good at making one type of identification (e.g., that a face pair is an identity match) are poor at another (e.g., that a face pair is a mismatch; Fysh & Bindemann, 2023; Megreya & Burton, 2007; Sauerland et al., 2016). This poses a great problem for developing a cognitive theory to explain how match and mismatch identifications are linked.

The current paper solves this puzzle. First, we have demonstrated the presence of systematic response biases in face matching, whereby observers tend to choose one response option (e.g., match) over the other (e.g., mismatch). These response biases reflect the decision-making thresholds that observers adopt to distinguish matches from mismatches. We provide evidence that the placement of these decision-making thresholds is stable at the individual level and transfers across tests, by demonstrating split-half reliability and correlations of criterion values. Importantly, we then demonstrate consistent positive correlations in match and mismatch performance when these criterion values are partialled out from the accuracy data. This phenomenon was observed with large sample sizes, totalling nearly 550 participants across two experiments that were conducted in two different labs, and three face-matching tests (AIFMT, GFMT, KFMT). This provides confidence that these are robust and replicable effects.

We then confirmed these findings with a series of simulations that systematically examined how variability in criterion influences the relationship between performance on match and mismatch trials in face-matching tasks or old and new trials in face-recognition tasks. The first of these simulations reproduced the dissociation between match and mismatch performance when criterion is free to vary, and the positive relationship between match and mismatch accuracy when criterion is controlled. The second simulation confirmed that the directionality of

the relationship between match and mismatch performance is influenced by the amount of variability in criterion and sensitivity. The third simulation, which represented that of an old/new recognition paradigm, because an item cannot be both a match and mismatch in an unfamiliar face matching task, then contrasted variability in criterion at the participant- and item-level. This showed that performance on old and new trials correlates tightly when both sources of variability are equivalent. In contrast, when one of these sources of variability is controlled while the other remains free to vary, a positive correlation between match and mismatch performance is only observed with the controlled parameter. This selective influence of criterion demonstrates directly that criterion determines the nature of the relationship in performance on old and new trials—a finding that also has implications for both the relationships between performance on old and new trials, and also on match and mismatch trials. Whereas recognition (match trials, old trials) and discrimination (mismatch trials, new trials) are associated strongly when criterion is controlled, variability in criterion disconnects these measures. Collectively, this can explain why match and mismatch performance and old/new performance is often uncorrelated in behavioural experiments (e.g., Fysh & Bindemann, 2023; Megreya & Burton, 2007; Megreya et al., 2011; Sauerland et al., 2016). This simulation also demonstrates why weak negative correlations between match and mismatch accuracy are often observed (e.g., Bate et al., 2018, 2019; Fysh & Bindemann, 2023; Kokje et al., 2018; Megreya & Burton, 2007). Whereas match and mismatch performance, as well as old and new performance, correlate positively and strongly when variability in criterion on a by-subjects or by-item level is eliminated systematically, reflecting a large contribution of variability in sensitivity, weak negative correlations are found when participants and items provide concurrent sources of variability. As individual differences in face matching and memory are shown for both accuracy or sensitivity and also criterion (e.g., Baker et al., 2023; Bobak et al., 2023; Fysh & Bindemann, 2018; McCaffery et al., 2018), and the difficulty of match and mismatch items is often free to vary across these trials, this can explain why weak negative correlations between match and mismatch accuracy are frequently found. Collectively, these findings demonstrate that individual differences in criterion placement mask the relationship in performance between match and mismatch performance, as well as in old/new face memory performance. In turn, these individual differences also provide the mechanism by which the identification of match and mismatch face pairings (or old and new recognition memory) is linked.

This possibility can be mapped onto a number of observations in the face-matching literature. For example, while there is evidence that both match and mismatch identifications reflect the perceived similarity of faces (Fysh & Bindemann, 2023; Papesch, 2018; Rice et al., 2013; Robertson et al., 2017), it has been suggested that these identifications might also require some separable deliberations. Match identifications might require a greater appreciation of the range of within-person variability that a person can exhibit in their appearance across different images or encounters, so that even two very different images of the same face can be reconciled as belonging to the same person. Mismatch identifications, might require a greater appreciation of whether similarities between people are meaningful or incidental. Accordingly, there is evidence that identity matches might be resolved through an accumulation of similarity information that two face images share, whereas mismatch identifications might rely more on an appraisal of meaningful similarities and differences between faces (see Bindemann & Burton, 2021; Fysh & Bindemann, 2023).

While similar reasoning might extend to the finding that some observers are better at identifying either match or mismatch trials, a second potential conclusion could be drawn. It is possible that enhanced accuracy with one of these stimulus types reflects differences in observer's decision-making thresholds. For example, two groups that are considered to be exceptionally skilled in face identification differ in their criterion placement and therefore show distinct error patterns. So-called super-recognizers are individuals who appear to have a high natural

ability to remember and identify faces (e.g., Nador et al., 2021; Ramon, 2021; Ramon et al., 2019). These observers exhibit a tendency to classify face pairs as matches and are biased to make highly confident errors on mismatch trials (Towler et al., 2023). Forensic facial examiners, on the other hand, are practitioners with a trained ability for face identification and tend to be more balanced in the placement of the decision-making criterion between matches and mismatches. These experts are also more cautious in expressing confidence when errors are made (Towler et al., 2023).

The current findings extend recent calls to integrate decision-making measures into models of face identification (Baker et al., 2023; Bindemann & Burton, 2021). Existing models focus primarily on the accuracy or sensitivity of identifications (e.g., Bruce & Young, 1986; Fitousi, 2023; O'Toole et al., 2018; Valentine, 1991), whereby errors occur because between-person similarity and within-person variability are hard to discriminate. For example, the unequal variance signal detection model (Fitousi, 2023) attributes the match-mismatch dissociation to independent signal distributions for these trial types that have their own variances, while correlations for both match and mismatch accuracy with the perceived similarity of face pairs is taken as evidence that these decisions also have a shared basis. Here we show that signal detection analysis can also provide insight into how *observer* characteristics influence face matching. Our findings extend contemporary theories by demonstrating that individual differences in the placement of criterion mask the overarching relationship in the identification of match and mismatch trials. While accuracy and sensitivity can determine the number of errors one makes, the placement of criterion determines the *type* of error. By taking these individual differences into account, the processing of matches and mismatches can be unified with a framework in which the placement of these decision-making thresholds provides the mechanism that links the identification of these face pairings.

Our study focused on criterion *placement*. A question that arises is whether criterion *shifts* might also create a correlation of match and mismatch performance by attenuating individual differences. Criterion shifts have been observed under a range of conditions, for example, when face matching is performed in an airport context (Bindemann et al., 2022), in response to costs, payoffs and unequal base rates (Baker & Bindemann, 2025; Stabile et al., 2024; for a discussion of these effects in other stimulus categories see Lynn & Barrett, 2014), and when superficial image differences (e.g., hue) are imposed on face pairs (Bobak et al., 2019a, 2019b). As our simulations show that a match-mismatch correlation emerges when criterion is controlled, criterion-shifting should exert similar effects if this leads to a reduction of individual differences in this parameter. While it is possible that criterion shifting might then be used as a strategy to reduce such individual differences, the success of such a strategy would be dependent on the shift in question driving all participants (or items) towards a neutral bias or in the same direction (for an example at the item-level, see Hilford et al., 2015). However, as there is evidence of individual differences in criterion placement at the participant level (e.g., Baker et al., 2023; Baker & Bindemann, 2025), it remains to be seen whether such an effect would be observed under conditions in which there are individual differences in performance.

Future research should certainly focus on the origin of individual differences in criterion placement in face matching. There is evidence that criterion is influenced by task difficulty and the relative frequency with which matches and mismatches occur (e.g., Baker & Bindemann, 2025; Stabile et al., 2024), as well as the applied context in which face matching is studied (e.g., Bindemann et al., 2022; Feng & Burton, 2019, 2021; McCaffery & Burton, 2016; Tummon et al., 2019). The key question, however, is what causes stable individual differences in criterion placement. While such an effect could arise from congenital differences in face identification ability (e.g., Wilmer et al., 2010; Zhu et al., 2010), a stronger argument might be made for the role of visual experience with faces. Experience with faces, for example, appears to shape performance on tasks that require memory for and the matching

of unfamiliar faces (Balas & Saville, 2015, 2017; Laurence & Mondloch, 2016; Baker et al., 2017; for similar experiential effects with face categories see Laurence et al., 2016; Proietti et al., 2019), and individual differences in face-matching ability (i.e., sensitivity) predict criterion shifting under variable conditions. For example, when there are unequal proportions of match and mismatch trials, or unequal costs associated with making different types of errors, observers with lower ability tend to adopt more extreme criteria towards the more likely or less costly response. Thus, it is possible that individual differences in criterion placement are less influential when ability is high, or even that individuals are less likely to strategically shift their criterion if the shift might require them to sacrifice accuracy (Miller & Kantner, 2020). Together, these studies suggest that individual differences in the visual experience with faces might also explain individual variation in criterion placement. Evidence in support of this possibility is found in studies that use faces from categories with which individuals have abundant versus limited experience, such as with own- versus other-race faces. Compared to participants who complete trials comprising only own-race faces, participants who complete trials comprising other-race faces show poorer performance, and when faced with unequal costs and base rates they show greater shifts in criterion (Stabile et al., 2024).

Finally, while the current study focused on unfamiliar face matching, whether similar effects also extend to *familiar* face recognition is an open question. We speculate, through the lens of Signal Detection Theory, that for familiar faces there is likely a large separation between signal and noise distributions in typical observers. This separation would be indicative of high levels of sensitivity and makes sense given reports of strong recognition performance for familiar faces (e.g., Jenkins et al., 2011; Zhou & Mondloch, 2016). This separation should render criterion placement mostly irrelevant—at least in what could be considered a relatively noise-free environment. This should yield similar effects to those reported by Megreya and Burton (2007), wherein participants showed a match-mismatch dissociation for unfamiliar faces, but performance correlated positively for familiarized faces. However, a second possibility to explain such an effect could come from the unequal variance signal detection model, which attributes the transition from unfamiliar face matching to familiar face recognition to a change in the similarity signal for a given face (Fitousi, 2023). Regardless of these possibilities, whether such a dissociation can be induced for familiar faces in a noisy environment (e.g., disguised familiar faces as in Noyes & Jenkins, 2019) is an intriguing question for future research, especially as representations of familiar faces are incredibly robust (e.g., Jenkins et al., 2011; Laurence et al., 2022; Zhou & Mondloch, 2016).

In conclusion, the experiments and simulations reported here demonstrate that the match-mismatch dissociation arises from individual differences in the placement of decision-making thresholds to distinguish these stimulus types. When these biases are controlled or partialled out from classification accuracy, reliable associations between match and mismatch identifications are found. These findings support a unitary cognitive theory in which individual differences in the placement of decision-making thresholds provide the mechanism by which the identification of face matches and mismatches are linked.

CRedit authorship contribution statement

Kristen A. Baker: Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Catherine J. Mondloch:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Peter J.B. Hancock:** Writing – review & editing, Methodology, Formal analysis. **Markus Bindemann:** Writing – review & editing, Supervision, Methodology, Data curation.

Acknowledgements

This research was funded by a Discovery Grant from the Natural

Sciences and Engineering Research Council of Canada (RGPIN 2022-04386) awarded to CJM, and a Newton International Fellowship awarded to KAB by the Royal Society (# 221754). We also thank the participants for their time in completing this research.

Data availability

Data is on OSF, and will be made public upon acceptance.

References

- Baker, K. A., & Bindemann, M. (2025). Decision-making framing in facial image comparison. *Journal of Applied Research in Memory and Cognition*. <https://doi.org/10.1037/mac0000216>. Advanced online publication.
- Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition*, 161, 19–30. <https://doi.org/10.1016/j.cognition.2016.12.012>
- Baker, K. A., & Mondloch, C. J. (2019). Two sides of face learning: Improving between-identity discrimination while tolerating more within-person variability in appearance. *Perception*, 48(11), 1124–1145. <https://doi.org/10.1177/0301006619867862>
- Baker, K. A., & Mondloch, C. J. (2022). Picture this: Photographers no better than controls for recognizing unfamiliar faces. *Perception*, 51(8), 591–595. <https://doi.org/10.1177/03010066221098727>
- Baker, K. A., & Mondloch, C. J. (2023). Unfamiliar face matching ability predicts the slope of face learning. *Scientific Reports*, 13(1), 5248. <https://doi.org/10.1038/s41598-023-32244-w>
- Baker, K. A., Stabile, V. J., & Mondloch, C. J. (2023). Stable individual differences in unfamiliar face identification: Evidence from simultaneous and sequential matching tasks. *Cognition*, 232, Article 105333. <https://doi.org/10.1016/j.cognition.2022.105333>
- Balas, B., & Saville, A. (2015). N170 face specificity and face memory depend on hometown size. *Neuropsychologia*, 69, 211–217. <https://doi.org/10.1016/j.neuropsychologia.2015.02.005>
- Balas, B., & Saville, A. (2017). Hometown size affects the processing of naturalistic face variability. *Vision Research*, 141, 228–236. <https://doi.org/10.1016/j.visres.2016.12.005>
- Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, 3(1), 1–19. <https://doi.org/10.1186/s41235-018-0116-5>
- Berger, A., Fry, R., Bobak, A. K., Juliano, A., & DeGutis, J. (2022). Distinct abilities associated with matching same identity faces versus discriminating different faces: Evidence from individual differences in prosopagnosics and controls. *Quarterly Journal of Experimental Psychology*, 75(12), 2256–2271. <https://doi.org/10.1177/17470218221076817>
- Bindemann, M., & Burton, A. M. (2021). Steps towards a cognitive theory of unfamiliar face matching. *Forensic Face Matching: Research and Practice*, 38–61.
- Bindemann, M., Fysh, M. C., Trifonova, I. V., Allen, J., McCall, C., & Burton, A. M. (2022). Face identification in the laboratory and in virtual worlds. *Journal of Applied Research in Memory and Cognition*, 11(1), 120. <https://doi.org/10.1016/j.jarmac.2021.07.010>
- Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, 40(5), 625–627. <https://doi.org/10.1068/p700>
- Bobak, A. K., Jones, A. L., Hilker, Z., Mestry, N., Bate, S., & Hancock, P. J. (2023). Data-driven studies in face identity processing rely on the quality of the tests and data sets. *Cortex*, 166, 348–364. <https://doi.org/10.1016/j.cortex.2023.05.018>
- Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019b). A grey area: How does image hue affect unfamiliar face matching? *Cognitive Research: Principles and Implications*, 4, 1–10. <https://doi.org/10.1186/s41235-019-0174-3>
- Bobak, A. K., Mileva, V. R., & Hancock, P. J. B. (2019a). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, 72(4), 872–881. <https://doi.org/10.1177/1747021818776145>
- Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, 5(4), 339. <https://doi.org/10.1037/1076-898X.5.4.339>
- Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, 7(3), 207. <https://doi.org/10.1037/1076-898X.7.3.207>
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3), 305–327. <https://doi.org/10.1111/j.2044-8295.1986.tb02199.x>
- Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, 42(1), 286–291. <https://doi.org/10.3758/BRM.42.1.286>
- Chen, W., Lander, K., & Liu, C. H. (2011). Matching faces with emotional expressions. *Frontiers in Psychology*, 2, 206. <https://doi.org/10.3389/fpsyg.2011.00206>
- Dale, G., & Arnell, K. M. (2015). Multiple measures of dispositional global/local bias predict attentional blink magnitude. *Psychological Research*, 79, 534–547. <https://doi.org/10.1007/s00426-014-0591-3>
- Dobbins, I. G., & Kroll, N. E. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6), 1186. <https://doi.org/10.1037/0278-7393.31.6.1186>
- Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44(4), 576–585. <https://doi.org/10.1016/j.neuropsychologia.2005.07.001>
- Duchaine, B. C., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, 24(4), 419–430.
- Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, 5(7), 589–601. <https://doi.org/10.1068/i0669>
- Estudillo, A. J., & Wong, H. K. (2022). Two face masks are better than one: Congruency effects in face matching. *Cognitive Research: Principles and Implications*, 7(1), 1–8. <https://doi.org/10.1186/s41235-022-00402-9>
- Favelle, S., Hill, H., & Claes, P. (2017). About face: Matching unfamiliar faces across rotations of view and lighting. *i-Perception*, 8(6). <https://doi.org/10.1177/2041669517744221>
- Feng, X., & Burton, A. M. (2019). Identity documents bias face matching. *Perception*, 48(12), 1163–1174. <https://doi.org/10.1177/0301006619877821>
- Feng, X., & Burton, A. M. (2021). Understanding the document bias in face matching. *Quarterly Journal of Experimental Psychology*, 74(11), 2019–2029. <https://doi.org/10.1177/17470218211017902>
- Fitousi, D. (2023). A signal detection–based confidence–similarity model of face matching. *Psychological Review*. <https://doi.org/10.1037/rev0000435>
- Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of Psychology*, 109(2), 219–231. <https://doi.org/10.1111/bjop.12260>
- Fysh, M. C., & Bindemann, M. (2023). Understanding face matching. *Quarterly Journal of Experimental Psychology*, 76(4), 862–880. <https://doi.org/10.1177/17470218221104476>
- Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1), 5–16. <https://doi.org/10.1037/0278-7393.16.1.5>
- Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, 100(3), 546. <https://doi.org/10.1037/0033-295X.100.3.546>
- Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, 16(3), 431–455. <https://doi.org/10.3758/PBR.16.3.431>
- Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition*, 69, 14–25. <https://doi.org/10.1016/j.concog.2019.01.008>
- Hancock, P. J. B. (2021). Familiar faces as islands of expertise. *Cognition*, 214, Article 104765. <https://doi.org/10.1016/j.cognition.2021.104765>
- Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996a). Face processing: Human perception and principal components analysis. *Memory & Cognition*, 24, 26–40. <https://doi.org/10.3758/BF03197270>
- Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996b). Face processing: Human perception and principal components analysis. *Memory and Cognition*, 24(1), 26–40.
- Hilford, A., Maloney, L. T., Glanzer, M., & Kim, K. (2015). Three regularities of recognition memory: The role of bias. *Psychonomic Bulletin & Review*, 22, 1646–1664. <https://doi.org/10.3758/s13423-015-0829-0>
- Hill, H., & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4), 986. <https://doi.org/10.1037/0096-1523.22.4.986>
- Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(2), 275. <https://doi.org/10.1037/0278-7393.20.2.275>
- Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, 121(3), 313–323. <https://doi.org/10.1016/j.cognition.2011.08.001>
- Kokje, E., Bindemann, M., & Megreya, A. M. (2018). Cross-race correlations in the abilities to match unfamiliar faces. *Acta Psychologica*, 185, 13–21. <https://doi.org/10.1016/j.actpsy.2018.01.006>
- Laurence, S., Baker, K. A., Proietti, V. M., & Mondloch, C. J. (2022). What happens to our representation of identity as familiar faces age? Evidence from priming and identity aftereffects. *British Journal of Psychology*, 113(3), 677–695. <https://doi.org/10.1111/bjop.12560>
- Laurence, S., & Mondloch, C. J. (2016). That's my teacher! Children's ability to recognize personally familiar and unfamiliar faces improves with age. *Journal of Experimental Child Psychology*, 143, 123–138. <https://doi.org/10.1016/j.jecp.2015.09.030>
- Laurence, S., Zhou, X., & Mondloch, C. J. (2016). The flip side of the other-race coin: They all look different to me. *British Journal of Psychology*, 107(2), 374–388. <https://doi.org/10.1111/bjop.12147>
- Liu, C. H., Chen, W., Han, H., & Shan, S. (2013). Effects of image preprocessing on face matching and recognition in human observers. *Applied Cognitive Psychology*, 27(6), 718–724. <https://doi.org/10.1002/acp.2967>
- Luong. (2025). Truncated Gaussian. <https://www.mathworks.com/matlabcentral/fileexchange/23832-truncated-gaussian>. MATLAB Central File Exchange. Retrieved March 26, 2025.
- Lynn, S. K., & Barrett, L. F. (2014). “Utilizing” signal detection theory. *Psychological Science*, 25(9), 1663–1673. <https://doi.org/10.1177/0956797614541991>
- Matthews, C. M., & Mondloch, C. J. (2018). Finding an unfamiliar face in a line-up: Viewing multiple images of the target is beneficial on target-present trials but costly

- on target-absent trials. *British Journal of Psychology*, 109(4), 758–776. <https://doi.org/10.1111/bjop.12301>
- McCaffery, J. M., & Burton, A. M. (2016). Passport checks: Interactions between matching faces and biographical details. *Applied Cognitive Psychology*, 30(6), 925–933. <https://doi.org/10.1002/acp.3281>
- McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, 3, 1–15. <https://doi.org/10.1186/s41235-018-0112-9>
- McIntyre, A. H., Hancock, P. J., Kittler, J., & Langton, S. R. (2013). Improving discrimination and face matching with caricature. *Applied Cognitive Psychology*, 27(6), 725–734. <https://doi.org/10.1002/acp.2966>
- Megreya, A. M., & Bindemann, M. (2015). Developmental improvement and age-related decline in unfamiliar face matching. *Perception*, 44(1), 5–22. <https://doi.org/10.1068/p7825>
- Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS One*, 13(3), Article e0193455. <https://doi.org/10.1371/journal.pone.0193455>
- Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, 69, 1175–1184. <https://doi.org/10.3758/BF03193954>
- Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, 27(6), 700–706. <https://doi.org/10.1002/acp.2965>
- Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology*, 64(8), 1473–1483. <https://doi.org/10.1080/17470218.2011.575228>
- Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception*, 44(11), 1332–1341. <https://doi.org/10.1177/0301006615599902>
- Mileva, M., Young, A. W., Jenkins, R., & Burton, A. M. (2020). Facial identity across the lifespan. *Cognitive Psychology*, 116, 101260. <https://doi.org/10.1016/j.cogpsych.2019.101260>
- Miller, M. B., & Kantner, J. (2020). Not all people are cut out for strategic criterion shifting. *Current Directions in Psychological Science*, 29(1), 9–15. <https://doi.org/10.1177/0963721419872747>
- Nador, J. D., Zoia, M., Pachai, M. V., & Ramon, M. (2021). Psychophysical profiles in super-recognizers. *Scientific Reports*, 11(1), 13184. <https://doi.org/10.1038/s41598-021-92549-6>
- Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition*, 165, 97–104. <https://doi.org/10.1016/j.cognition.2017.05.012>
- Noyes, E., & Jenkins, R. (2019). Deliberate disguise in face identification. *Journal of Experimental Psychology: Applied*, 25(2), 280. <https://doi.org/10.1037/xap0000213>
- O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, 22(9), 794–809. <https://doi.org/10.1016/j.tics.2018.06.006>
- Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications*, 3(1), 1–9. <https://doi.org/10.1186/s41235-018-0110-y>
- Papesh, M. H., Heisick, L. L., & Warner, K. A. (2018). The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied*, 24(3), 416. <https://doi.org/10.1037/xap0000156>
- Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., ... Weimer, S. (2011). An introduction to the good, the bad, & the ugly face recognition challenge problem. In: 2011 IEEE international conference on Automatic Face & Gesture Recognition (FG) (pp. 346–353). IEEE. <https://doi.org/10.1109/FG.2011.5771424>
- Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1), 74–85. <https://doi.org/10.1016/j.imavis.2013.12.002>
- Proietti, V., Laurence, S., Matthews, C. M., Zhou, X., & Mondloch, C. J. (2019). Attending to identity cues reduces the own-age but not the own-race recognition advantage. *Vision Research*, 157, 184–191. <https://doi.org/10.1016/j.visres.2017.11.010>
- Ramon, M. (2021). Super-recognizers—a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, 158, Article 107809. <https://doi.org/10.1016/j.neuropsychologia.2021.107809>
- Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, 110(3), 461–479. <https://doi.org/10.1111/bjop.12368>
- Rice, A., Phillips, P. J., & O'Toole, A. (2013). The role of the face and body in unfamiliar person identification. *Applied Cognitive Psychology*, 27(6), 761–768. <https://doi.org/10.1002/acp.2969>
- Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, 70(5), 897–905. <https://doi.org/10.1080/17470218.2015.1136656>
- Robertson, D. J., Kramer, R. S., & Burton, A. M. (2017). Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PLoS One*, 12(3), Article e0173319. <https://doi.org/10.1371/journal.pone.0173319>
- Sauerland, M., Sagana, A., Siegmann, K., Heiligers, D., Merckelbach, H., & Jenkins, R. (2016). These two are different. Yes, they're the same: Choice blindness for facial identity. *Consciousness and Cognition*, 40, 93–104. <https://doi.org/10.1016/j.concog.2016.01.003>
- Stabile, V. J., Baker, K. A., & Mondloch, C. J. (2024). Criterion shifting in an unfamiliar face-matching task: Effects of base rates, payoffs, and perceptual discriminability. *Journal of Applied Research in Memory and Cognition*. <https://doi.org/10.1037/mac0000157>
- Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. *Cognitive Research: Principles and Implications*, 5, 1–17. <https://doi.org/10.1186/s41235-019-0205-0>
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. <https://doi.org/10.3758/BF03207704>
- Strathie, A., Hughes-White, N., & Laurence, S. (2022). The sibling familiarity effect: Is within-person facial variability shared across siblings? *British Journal of Psychology*, 113(1), 327–345. <https://doi.org/10.1111/bjop.12517>
- Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6), 1379. <https://doi.org/10.1037/0278-7393.24.6.1379>
- Summerfield, C., & Egner, T. (2014). Attention and decision-making. In *The Oxford handbook of decision making*. Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780199675111.013.018>
- Towler, A., Dunn, J. D., Castro Martínez, S., Moreton, R., Eklöf, F., Ruifrok, A., & White, D. (2023). Diverse types of expertise in facial recognition. *Scientific Reports*, 13(1), 11396. <https://doi.org/10.1038/s41598-023-28632-x>
- Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS One*, 14(2), Article e0211037. <https://doi.org/10.1371/journal.pone.0211037>
- Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (2021). Diagnostic feature training improves face matching accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(8), 1288. <https://doi.org/10.1037/xlm0000972>
- Tummon, H. M., Allen, J., & Bindemann, M. (2019). Facial identification at a virtual reality airport. *i-Perception*, 10(4), Article 2041669519863077. <https://doi.org/10.1177/2041669519863077>
- Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, 43(2), 161–204. <https://doi.org/10.1080/14640749108400966>
- Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, 20, 291–302. <https://doi.org/10.3758/BF03199666>
- White, D., & Burton, A. M. (2022). Individual differences and the multidimensional nature of face perception. *Nature Reviews Psychology*, 1(5), 287–300. <https://doi.org/10.1038/s44159-022-00041-3>
- White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS One*, 10(10), Article e0139827. <https://doi.org/10.1371/journal.pone.0139827>
- White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS One*, 9(8), Article e103510. <https://doi.org/10.1371/journal.pone.0103510>
- Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, 107(11), 5238–5241. <https://doi.org/10.1073/pnas.0913053107>
- Wirth, B. E., & Carbon, C. C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*, 23(2), 138. <https://doi.org/10.1037/xap0000114>
- Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(4), 681. <https://doi.org/10.1037/0278-7393.18.4.681>
- Yang, T., Penton, T., Köybaşı, Ş. L., & Banissy, M. J. (2017). Social perception and aging: The relationship between aging and the perception of subtle changes in facial happiness and identity. *Acta Psychologica*, 179, 23–29. <https://doi.org/10.1016/j.actpsy.2017.06.006>
- Zhou, X., Matthews, C. M., Baker, K. A., & Mondloch, C. J. (2018). Becoming familiar with a newly encountered face: Evidence of an own-race advantage. *Perception*, 47(8), 807–820. <https://doi.org/10.1177/0301006618783915>
- Zhou, X., & Mondloch, C. J. (2016). Recognizing “Bella Swan” and “Hermione Granger”: No own-race advantage in recognizing photos of famous faces. *Perception*, 45(12), 1426–1429. <https://doi.org/10.1177/030100661666204>
- Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, 20(2), 137–142. <https://doi.org/10.1016/j.cub.2009.11.067>