Individual differences in face and voice matching abilities: The relationship between accuracy and consistency

Robin S. S. Kramer[1], Alex L. Jones[2], & Georgina Gous[1]

[1]School of Psychology, University of Lincoln, Lincoln, UK

[2]Department of Psychology, Swansea University, Swansea, UK

Corresponding Author:

Robin Kramer, School of Psychology, University of Lincoln, Lincoln LN6 7TS, UK.

E-mail: remarknibor@gmail.com

Telephone: +44 (0)1522 835806

Running head: Individual differences in face and voice matching abilities

Conflicts of interest

There are no conflicts of interest to be disclosed.

Data availability statement

The data that support the findings of these experiments are available through the Open Science Framework at https://osf.io/8gtue/?view_only=8ee658e2ad8040adb1977b386c17da9d.

Acknowledgements

The authors thank their Research Skills IV students, along with Abi Davis, for recruiting participants in both experiments.

Abstract

Deciding whether two different face photographs or voice samples are from the same person represent fundamental challenges within applied settings. To date, most research has focussed on average performance in these tests, failing to consider individual differences and within-person consistency in responses. Here, participants completed the same face (Experiment 1) or voice matching test (Experiment 2) on two separate occasions, allowing comparison of overall accuracy across the two timepoints as well as consistency in trial-level responses. In both experiments, participants were highly consistent in their performances. In addition, we demonstrated a large association between consistency and accuracy, with the most accurate participants also tending to be the most consistent. This is an important result for applied settings in which organisational groups of super-matchers are deployed in real-world contexts. Being able to reliably identify these high performers based upon only a single test informs regarding recruitment for law enforcement agencies worldwide.

Keywords

Face matching; voice matching; individual differences; consistency; accuracy

1 Introduction

Deciding whether two different face photographs are of the same person, or whether a person standing in front of you is the same person depicted in a photograph, represent fundamental challenges within applied identification settings. For example, passport renewal typically involves the former while border force officers are regularly faced with the latter. Across a variety of contexts and designs, researchers have demonstrated that such tasks are error-prone (e.g., Bruce et al., 1999, 2001; Kemp, Towell, & Pike, 1997; Megreya & Burton, 2006, 2008). Although numerous studies have focussed on accuracy in what has been termed 'face matching', there is little investigation of how consistent individuals are in their performance.

The difficulty in carrying out face matching tasks lies in our limited ability to cope with unfamiliar faces. In general, the problem is one of "telling faces together" – realising that highly dissimilar images may still depict the same person (Jenkins, White, Van Montfort, & Burton, 2011). While familiar facial comparisons are simple to perform, consistently resulting in ceiling-level accuracies (e.g., Bruce et al., 2001), we are significantly worse with the faces of unfamiliar people because such decisions are closely bound to the visual properties of the particular images (Hancock, Bruce, & Burton, 2000), resulting in a qualitatively different process (less reliance on configural processing and more comparable with inverted faces – Megreya & Burton, 2006). Even small changes in lighting, viewpoint, facial expression, and distance to camera, for example, can have substantial negative effects on matching accuracy (e.g., Estudillo & Bindemann, 2014; Noyes & Jenkins, 2017). Interestingly, metacognitive research suggests that people are somewhat blind to this familiarity distinction, incorrectly predicting that the faces they themselves are familiar with will be more accurately matched by unfamiliar others (Ritchie et al., 2015).

While the majority of studies have focussed on establishing the difficulties that people *in general* demonstrate when performing face matching in a variety of situations (e.g., Kramer, Mohamed, & Hardy, 2019; Kramer, Mulgrew, & Reynolds, 2018; Ritchie et al., 2018; White, Burton, Jenkins, & Kemp, 2014), others have begun to identify and explore the substantial individual differences that are apparent (Burton, White, & McNeill, 2010; Fysh, 2018; Fysh & Bindemann, 2018; McCaffery, Robertson, Young, & Burton, 2018; Stacchi, Huguenin-Elie, Caldara, & Ramon, 2020; White, Kemp, Jenkins, Matheson, & Burton, 2014; for a review, see Lander, Bruce, & Bindemann, 2018). Initial focus on face recognition has identified a large range of abilities, with prosopagnosics (McConachie, 1976) and super-recognisers (Russell, Duchaine, & Nakayama, 2009) featuring at the extremes of this natural continuum. Similarly, considerable differences across individuals have been found in research on face matching (e.g., Burton et al., 2010). Having established this between-person variability in performance on such tasks, it follows that researchers next considered the nature of within-person variability. In other words, how consistent is an individual in their performance on these tasks?

One way to approach consistency is to investigate performance across different tasks involving face processing. For example, researchers interested in the scope of super-recognisers' abilities have considered how they perform in both recognition and matching domains. Evidence supports the idea that this particular group excels in both tasks in some studies, while other work has suggested that super-recognisers and super-matchers may represent distinct samples of people (Bate et al., 2019; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; Davis, Lander, Evans, & Jansari, 2016; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). In the general population, face matching ability is associated with performance on face memory tasks (McCaffery et al., 2018; Verhallen et al., 2017), as well as those involving searching for faces in crowds (Kramer, Hardy, & Ritchie,

2020). Overall, these findings imply that there may be some generic, underlying ability with faces that results in good performance across all tests (e.g., the factor *f* – Verhallen et al., 2017). This has real-world importance if, for example, individuals are recruited with the goal that they perform at a high level in a variety of face-related tasks.

The second approach to considering an individual's consistency is through their repeated performance on the same task. If we are to characterise people as 'good' or 'bad' at face matching, for instance, then we necessarily require that they perform consistently when faced with a particular test. By using only a single measure of performance, an individual's ability is confounded with a variety of other factors particular to the time of testing, such as fatigue, illness, lifestyle influences (e.g., alcoholic drinking), and perhaps most importantly, luck. Typically, there is an element of guessing when completing any test, and so accuracy may be over- or under-estimated when limited to this single timepoint. By investigating repeat performances, we are able to 1) provide a more precise and reliable estimate of ability; and 2) determine whether measures of accuracy and consistency are associated.

For face recognition, evidence suggests that abilities demonstrate a strong genetic basis (Shakeshaft & Plomin, 2015; Wilmer et al., 2010), providing a reason to expect stable performance across time within individuals. Although such research has yet to be undertaken with regard to face matching, other types of studies have begun to investigate this topic. For example, when participants were asked to complete a similar face matching test on each of five consecutive days,[1] researchers identified considerable variability in performance both between and within individuals (Experiment 2 – Bindemann, Avetisyan, & Rakow, 2012). Further, after removing the most accurate performers (those who averaged over 97% and were necessarily also highly consistent, given their near-ceiling performances), the authors

[1] Forty trials per day, selected from an initial set of 200 face pairs so that each subtest was matched in terms of difficulty.

found no significant association between accuracy and consistency. However, more recently, Bate and colleagues (2019) investigated the performance of police officers (previously identified as proficient on face-related tasks) across three blocks of a matching test, where face pairs differed in either pose, the presence of glasses, or facial hair. When considering performance on match and mismatch trials separately, these researchers found large associations between accuracy and consistency for both types of trial: individuals who were more accurate across these three blocks were also more consistent.

While these studies investigated performance by the same participants across similar tests of face matching, to our knowledge, only one study has investigated repeated performance on the same test. By asking participants to complete a 200-trial face matching test on each of three consecutive days, the researchers were able to identify errors in which trials were responded to correctly on one day but incorrectly on the next, and vice versa (Experiment 1 – Bindemann et al., 2012). Interestingly, participants were equally as likely to produce these two types of errors. In addition, after removing the most accurate performers (see above), no significant association was found between the accuracy and consistency of participants (although the authors did not consider match and mismatch trials separately for this analysis).

Taken together, research to date appears to provide mixed results regarding consistency. On the one hand, there is evidence to suggest that there may be an underlying ability with face processing that means people can be categorised as ‘good’ or ‘bad’ with faces in general, irrespective of the task (although any associations appear to be far from perfect). On the other hand, substantial within-person inconsistency has been identified both across similar tasks and within the same task across testing days, arguing that individuals can often be inconsistent in their performance. Given that the goal from an applied perspective is to identify and recruit those who perform both accurately and consistently, the evidence is

again unclear at present. Studies have found both the presence (Bate et al., 2019) and absence (Bindemann et al., 2012) of an association between accuracy and consistency.

While the focus within the literature has been clearly placed upon applied contexts in which faces are used for identification, recent research has begun to investigate similar questions as they relate to voice processing. Voice matching abilities are important within forensic contexts in which perpetrators are encountered under poor visual conditions or when an offence is committed over the telephone. Previous research has identified large individual differences in voice matching and identification abilities (Lavan, Burston, & Garrido, 2019; Mühl, Sheil, Jarutytė, & Bestelmeyer, 2018) and researchers have begun to investigate the possibility that individuals may be classified as super-voice-recognisers (Jenkins et al., 2020). Evidence also suggests that performance on face and voice matching tasks are weakly correlated (Jenkins et al., 2020; Mühl et al., 2018), perhaps suggesting the presence of a cross-modality mechanism underlying person perception more generally. Irrespective of whether performance across the two modalities is related, it is important to consider, as with faces, the nature of accuracy versus consistency within voice matching. As outlined above, this is best examined through individuals completing the same test more than once.

In the current experiments, we therefore aim to investigate the relationship between an individual's accuracy and consistency. While previous work has utilised matching tasks involving binary 'same'/'different' responses (Bate et al., 2019; Bindemann et al., 2012), here we incorporate a rating scale in order to allow more fine-grained analyses regarding the comparison between responses across timepoints. Although this can be achieved through the addition of a confidence rating alongside participants' binary responses (e.g., White, Burton, et al., 2014), we instead utilise a response scale incorporating both the decision and its associated confidence (O'Toole et al., 2007) in order to allow for a simple, correlational approach when comparing responses given over the two sessions. Further, we consider

repeated testing on the same task but opt for a minimum of one week between sessions, compared with consecutive days used previously (Bindemann et al., 2012). In this way, we can be more confident that participants are unable to remember and reproduce responses across sessions, but also that specific factors (e.g., fatigue or illness) will not remain constant. Finally, we extend the investigation of accuracy and consistency beyond facial images to include voice matching for the first time.

2 Experiment 1

In this first experiment, face matching was assessed across two testing sessions which were separated by at least one week. In both sessions, the same 40 face pairs were presented. This design allowed us to investigate both accuracy and consistency, as well as the potential for an association between these measures.

2.1 Method

2.1.1 Participants

An initial sample of 67 volunteers participated in the first session of the experiment. However, due to attrition, only 50 individuals (age M = 21.1 years, SD = 4.9 years; 68% women; 86% self-reported as White) completed both sessions. All volunteers gave informed, onscreen consent before participating in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited through 'word of mouth' in order to allow the experimenters to keep track of who took part and to enable subsequent contact for participation in the second session.

The sample sizes for both experiments presented here were based on the number of participants used in earlier studies investigating the relationship between accuracy and consistency in face matching performance (30 participants – Bindemann et al., 2012). The reported correlation between these two factors (0.36; Bindemann et al., 2012) meant that a sample size of 55 was required after choosing an α of .05 and with power (1 - β) set to 0.80 (GPower 3.1 software; Faul, Erdfelder, Lang, & Buchner, 2007). As such, our target was at least 55 participants, although our final sample size was slightly below this due to attrition. Importantly however, it may not be sensible to base our power calculations on a single prior study, which was itself underpowered (0.53). Instead, with the final sample sizes presented here and in Experiment 2, we are able to say that the minimal correlation detectable was 0.39 (which is very similar to the effect reported in Bindemann et al., 2012).

Both experiments reported here were approved by the University of Lincoln's School of Psychology ethics committee (PSY1920106) and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

2.1.2 Stimuli

We used the short version of the Glasgow face matching test (GFMT; Burton et al., 2010) to assess performance. The task comprised 40 pairs of adult male (24) and female faces (16), where half the pairs were match trials (different images of the same person, taken approximately 15 minutes apart with different cameras) and half were mismatch trials (different people with a similar appearance). All images were greyscale, passport-style photographs, depicting a front-on, neutral expression, and displayed on a plain, white background (see Figure 1). The 40 face pairings were taken from the original GFMT set of 168 pairs and represented the most difficult trials (based on the performance of 300

participants; Burton et al., 2010). As such, the more challenging version of this test was chosen in order to minimise the risk of ceiling-level performances, which would necessarily affect our consideration of consistency (see below).

2.1.3 Procedure

The experiment was completed online using the Qualtrics survey platform (www.qualtrics.com). After consent was obtained, participants provided demographic information (age, sex, and ethnicity). In addition, in order to be able to identify the same individual's responses across the two sessions, we collected each participant's first name and date of birth. After data collection was completed and data files had been amalgamated, these pieces of information were deleted and all data were anonymised.

In the first experimental session, participants completed all 40 trials of the short version of the GFMT online. On each trial, two face photographs were displayed onscreen and participants were instructed to decide whether they thought these faces were the same person or different people. Following O'Toole et al. (2007), responses were provided using a labelled rating scale: 1) sure they are the same; 2) think they are the same; 3) don't know; 4) think they are not the same; 5) sure they are not the same. Trial order was randomised for each participant, no time limits were imposed upon responses, and no feedback was given at any stage.

All participants who completed the first session were contacted a week after they had taken part and were provided with a weblink for the second session. This online task was identical to the first session, with trial order again randomised for each participant.

Through the information provided onscreen before both sessions began, participants were informed that we were investigating consistency and that they would be completing the same task on two separate occasions.

2.2 Results

First, we calculated the interval between completing the two sessions for each participant. In all cases, at least one week had passed between sessions: M = 21.6 days, SD = 5.6 days; range = 12 to 35 days.

A summary of our main performance measures can be found in Table 1.

2.2.1 Conversion to Binary Responses

In order to be able to compare our results with those of Bindemann et al. (2012), we converted our data to binary responses. In line with previous work, responses 1 and 2 were deemed "same" judgements and responses 3, 4, and 5 were deemed "different" judgements (O'Toole et al., 2007).[2]

First, we considered whether participant accuracy (proportion correct) differed across testing sessions. However, a 2 (Session: T1, T2) x 2 (Trial Type: match, mismatch) within-subjects analysis of variance (ANOVA) found no significant main effect of Session, $F(1, 49) = 0.73$, $p = .397$, $\eta_p^2 = .01$, 90% CI [.00, .11], or Trial Type, $F(1, 49) = 0.56$, $p = .458$, $\eta_p^2 = .01$, 90% CI [.00, .10], and no interaction between these factors, $F(1, 49) = 0.04$, $p = .852$, $\eta_p^2 < .01$, 90% CI [.00, .04]. Therefore, in line with Bindemann et al. (2012), we found no overall difference in accuracies across the two sessions. Mean performance across all conditions was

[2] For analyses where a response of 3 was deemed "same", see the supporting information.

0.87, which was comparable with normative accuracy data on this test (M = 0.81; Burton et al., 2010).

Next, we explored the types of errors that participants made by examining the proportion of trials that received correct responses in both sessions (CC), correct responses in the first session and incorrect responses in the second (CE), incorrect responses in the first session and correct responses in the second (EC), and incorrect responses in both sessions (EE). These data are summarised in Figure 2. Combining CC and EE proportions produced a measure of consistency in participants' responses, while combining CE and EC represented inconsistency. Given that these two values were the complement of each other (necessarily summing to 1), we compared one of these error types to a value of 0.5 (since both comparisons would produce the same statistical result). Across all trials, we found significantly more consistency than inconsistency in responses, $t(49) = 20.10$, $p < .001$, Cohen's $d = 2.84$, 95% CI [2.21, 3.47]. Indeed, participants were consistent on 81.3% of trials on average. This greater consistency was also apparent when match, $t(49) = 13.69$, $p < .001$, Cohen's $d = 1.94$, 95% CI [1.46, 2.40], and mismatch trials, $t(49) = 15.25$, $p < .001$, Cohen's $d = 2.16$, 95% CI [1.64, 2.66], were analysed separately.

In addition to analysing consistency in trial-level responses across the two sessions, we investigated consistency at the level of the participant, i.e., in task-level performances. If participants were consistent across the two testing sessions, this result should be reflected in an association between their overall T1 and T2 accuracies. For combined performance (irrespective of trial type), we found a large association between T1 and T2 accuracies, $r_s(48) = .65$, $p < .001$, 95% CI [.45, .79]. Separate analyses of match, $r_s(48) = .51$, $p < .001$, 95% CI [.27, .69], and mismatch trials, $r_s(48) = .49$, $p < .001$, 95% CI [.25, .68], also produced this pattern of results.[3] Taken together, these findings demonstrated that face-matching

[3] We report Spearman's rank correlations here and below where one or both variables failed tests of normality.

performance was largely consistent across the two sessions, in general agreement with the results of Bate et al. (2019).

Task-level performance was also investigated using signal detection measures. We calculated sensitivity indices (*d'*) and response biases (*c*) using the following: *Hit* – both images were of the same identity and participants responded "same"; *False alarm* – the two images were of different people and participants responded "same". We found significant associations between T1 and T2 performances for both *d'*, $r_s(48) = .65$, $p < .001$, 95% CI [.45, .79], and *c* values, $r_s(48) = .42$, $p = .002$, 95% CI [.16, .63].

Finally, we investigated whether participants showed any relationship between accuracy and consistency in their task-level performances. Following Bindemann et al. (2012), accuracy here was calculated as the combined performance (proportion correct) for match and mismatch trials, averaged across the two sessions. Given that our participants only completed two sessions, inconsistency was simply calculated as the unsigned difference between the two performances (since larger values reflected greater inconsistency). Across all participants, we found a medium-sized association between accuracy and inconsistency, $r_s(48) = -.41$, $p = .003$, 95% CI [-.62, -.15]. This relationship was also found when match, $r_s(48) = -.51$, $p < .001$, 95% CI [-.69, -.27], and mismatch trials, $r_s(48) = -.58$, $p < .001$, 95% CI [-.74, -.36], were analysed separately. Following Bindemann et al. (2012), we also recalculated our combined value after excluding participants ($n = 4$) whose accuracies were greater than or equal to 0.97, given that those who performed near ceiling in both sessions would inevitably produce near-floor inconsistencies, potentially producing a statistically artificial result. Again, we found a medium-sized association between accuracy and inconsistency, $r_s(44) = -.32$, $p = .028$, 95% CI [-.56, -.03]. While Bindemann et al. (2012) suggested that these two measures were "broadly separable indices of face-matching ability"

(p. 283), we found that participants who were more accurate also tended to be more consistent.

2.2.2 Analysis of Ratings

Although we initially converted our ratings data to binary responses in order to compare our results directly with those of Bindemann et al. (2012), we also analysed the original ratings, providing additional insights concerning accuracy and consistency based on more fine-grained information regarding participants' perceptions. Rather than making explicit judgements about whether face pairs were 'same' or 'different', participants rated the likelihood that two images were of the same person (e.g., O'Toole et al., 2007). This approach allows representational and decisional components to be separated, which can be of benefit in certain contexts. For example, a system administrator can subsequently control the decisional criterion by manipulating gain according to the risk associated with specific types of error. For instance, if it is desirable to avoid 'miss' decisions (match trials given a 'different' response) then the threshold for 'same' responses could be set lower than if the priority is to avoid 'false alarms'.

For each participant, separately for each testing session, we calculated the hit and false alarm rates for each possible threshold along the rating scale (1 through 5). Plotting these values produced the receiver operating characteristic (ROC), with the area under this ROC curve (AUC) representing a measure that is widely used to assess the performance of classification rules over the entire range of possible thresholds (Krzanowski & Hand, 2009). As such, AUC allowed us to quantify the performance of a classifier (here, our participants), irrespective of where the cut-off between binary 'same'/'different' responses might have

been placed. Here, we used AUC to quantify the extent to which ratings discriminated between match and mismatch trials (e.g., White, Burton, Kemp, & Jenkins, 2013).

First, we considered whether participant classification performance (AUC) differed across testing sessions. However, a paired-samples *t*-test found no significant difference, $t(49) = 0.86$, $p = .394$, Cohen's $d = 0.12$, 95% CI [-.14, 0.37]. The mean AUC across both sessions was 0.91. Using these values, we also investigated task-level consistency by correlating AUC values for the two sessions. We found a large association between T1 and T2 performance, $r_s(48) = .65$, $p < .001$, 95% CI [.45, .79].

Next, we considered the consistency of participants' responses across the two sessions (i.e., at the trial level). To quantify this, we correlated each participant's T1 and T2 responses, with these values across participants demonstrating substantial within-person agreement, mean $r_s = .69$, 95% CI [.63, .75] (after applying Fisher's *r*-to-*z* transformation and its inverse as necessary).

Finally, we again investigated whether participants showed any relationship between accuracy and consistency in their task-level performances. Here, accuracy was calculated by averaging the AUC values for the two sessions, whereas inconsistency was calculated as the unsigned difference between the two values (with larger values reflecting greater inconsistency; see above). Across all participants, we found a large association between accuracy and inconsistency, $r_s(48) = -.56$, $p < .001$, 95% CI [-.73, -.33]. As above, we also recalculated this value after excluding the most accurate participants ($n = 4$) identified previously. Again, we found an association between accuracy and inconsistency, $r_s(44) = -.44$, $p = .002$, 95% CI [-.65, -.17]. Therefore, analysis of the original responses replicated our earlier finding that participants who were more accurate also tended to be more consistent.

2.2.3 Testing Interval and Consistency

Given the range in the number of days (12 to 35) between testing sessions, we considered whether this amount of time was associated with task-level consistency. Simply, participants may have been more consistent across sessions if less time had passed. However, we found no relationship between testing interval and consistency, whether quantified as the unsigned difference between T1 and T2 accuracies using proportion correct, $r_s(48) = -.09$, $p = .523$, 95% CI [-.36, .19], or AUC values, $r_s(48) = -.12$, $p = .404$, 95% CI [-.39, .16].

2.3 Discussion

This experiment investigated both consistency and accuracy using the short version of the GFMT, a benchmark test of unfamiliar face matching (Burton et al., 2010). In previous research, participants were asked to complete a similar test (200 trials using stimuli from the same database) on three consecutive days (Experiment 1 – Bindemann et al., 2012). Here, we incorporated two differences worth highlighting. First, our participants responded on each trial using a 1 to 5 scale, allowing additional analyses beyond those available following binary match/mismatch responses. Second, the interval between sessions was a minimum of 12 days, decreasing the likelihood that responses would be remembered and reproduced at T2.

In line with Bindemann et al. (2012), we found no overall differences in performance across our two sessions. However, in contrast with their work, we find strong support for consistency in performance. Whether considering participant accuracy as derived from binary data or the association between raw scale responses, we see substantial agreement across our two sessions. That is, at the trial level, participants tended to respond in the same way to the same stimulus pairs over time. In addition, we demonstrate a significant association at the

participant level between accuracy and consistency, which is medium-sized with our binary data and large when analysing our scale responses. As such, we argue that participants who perform more accurately also tend to be those who are more consistent, suggesting that recruitment strategies based upon accuracy alone may be justified.

3 Experiment 2

Experiment 1 demonstrated that participants responded in a largely consistent manner when completing the same task on two separate occasions. In addition, we found strong support for the idea that accuracy and consistency are associated. Importantly, these results apply to a face matching test and researchers have yet to consider these factors in relation to voice matching.

In this experiment, voice matching was assessed across two testing sessions which were separated by at least one week. In both sessions, the same 80 voice pairs were presented. This design allowed us to investigate both accuracy and consistency, as well as the potential for an association between these measures.

3.1 Method

3.1.1 Participants

An initial sample of 77 volunteers participated in the first session of the experiment. However, due to attrition, only 45 individuals (age M = 25.0 years, SD = 12.3 years; 69% women; 98% self-reported as White) completed both sessions. All volunteers gave informed, onscreen consent before participating in the experiment and were provided with an onscreen

debriefing upon completion. Participants were recruited using the same method as in Experiment 1. There was no overlap between this participant sample and those who took part in the previous experiment.

Data from one additional participant were excluded because inspection of their responses revealed that they had given the same rating on almost all trials in both sessions, resulting in overall accuracies of 51% and 48% for T1 and T2 respectively (where 50% was chance performance).

3.1.2 Stimuli

We used the Bangor voice matching test (BVMT; Mühl et al., 2018) to assess performance. The task comprised 80 pairs of adult male (40) and female voices (40), where half the pairs were match trials (different voice samples produced by the same person) and half were mismatch trials (voice samples produced by two different people). Each sample was either a consonant-vowel-consonant (e.g., "had") or vowel-consonant-vowel (e.g., "aba"). The 80 voice pairings were taken from an initial set of 288 pairs collected by the researchers and were chosen to span a wide range of difficulties (based on the performance of 457 participants; Mühl et al., 2018).

3.1.3 Procedure

The experiment was completed online using the Gorilla experiment builder (gorilla.sc). As in Experiment 1, we collected information regarding the participant's age, sex, and ethnicity, as well as their first name and date of birth.

In the first experimental session, participants completed all 80 trials of the BVMT online. On each trial, two buttons were displayed onscreen (labelled 'Play Sound 1' and 'Play Sound 2') and participants were instructed to decide whether they thought these samples were the same speaker or different speakers. As in Experiment 1, responses were provided using a 1-5 rating scale. Participants were able to listen to the voice samples an unlimited number of times by clicking on the two buttons, prior to giving their response. Between trials, a fixation cross appeared for 800 ms. Trial order was randomised for each participant, no time limits were imposed upon responses, and no feedback was given at any stage.

All participants who completed the first session were contacted a week after they had taken part and were provided with a weblink for the second session. This online task was identical to the first session, with trial order again randomised for each participant.

Through the information provided onscreen before both sessions began, participants were informed that we were investigating consistency and that they would be completing the same task on two separate occasions.

3.2 Results

First, we calculated the interval between completing the two sessions for each participant. In all cases, at least one week had passed between sessions: M = 21.0 days, SD = 6.7 days; range = 11 to 34 days.

A summary of our main performance measures can be found in Table 2.

3.2.1 Conversion to Binary Responses

Following the analyses carried out in our first experiment, responses 1 and 2 were deemed "same" judgements and responses 3, 4, and 5 were deemed "different" judgements (O'Toole et al., 2007).[4]

First, we considered whether participant accuracy (proportion correct) differed across testing sessions. However, a 2 (Session: T1, T2) x 2 (Trial Type: match, mismatch) within-subjects ANOVA found no significant main effect of Session, $F(1, 44) = 1.41$, $p = .241$, $\eta_p^2 = .03$, 90% CI [.00, .15], and no interaction between these factors, $F(1, 44) = 1.18$, $p = .284$, $\eta_p^2 = .03$, 90% CI [.00, .14]. A significant main effect of Trial Type, $F(1, 44) = 21.77$, $p < .001$, $\eta_p^2 = .33$, 90% CI [.15, .48], revealed that accuracy was higher for match trials ($M = 0.85$) in comparison with mismatch trials ($M = 0.71$). Therefore, we found no overall difference in accuracies across the two sessions. Mean performance across all conditions was 0.78, which was comparable with normative accuracy data on this test ($M = 0.85$; Mühl et al., 2018).

Next, we explored the types of errors that participants made by examining the proportion of CC, CE, EC, and EE responses. These data are summarised in Figure 3. Combining CC and EE proportions produced a measure of consistency in participants' responses, and we compared this to a value of 0.5. Across all trials, we found significantly more consistency than inconsistency in responses, $t(44) = 12.83$, $p < .001$, Cohen's $d = 1.91$, 95% CI [1.41, 2.40]. Indeed, participants were consistent on 73.7% of trials on average. This greater consistency was also apparent when match, $t(44) = 13.83$, $p < .001$, Cohen's $d = 2.06$, 95% CI [1.54, 2.58], and mismatch trials, $t(44) = 7.96$, $p < .001$, Cohen's $d = 1.19$, 95% CI [0.80, 1.56], were analysed separately.

In addition to analysing consistency in trial-level responses across the two sessions, we investigated consistency at the level of the participant, i.e., in task-level performances. If participants were consistent across the two testing sessions, this result should be reflected in

[4] For analyses where a response of 3 was deemed "same", see the supporting information.

an association between their overall T1 and T2 accuracies. For combined performance (irrespective of trial type), we found a large association between T1 and T2 accuracies, $r_s(43) = .65$, $p < .001$, 95% CI [.44, .79]. Separate analyses of match, $r_s(43) = .49$, $p = .001$, 95% CI [.23, .69], and mismatch trials, $r_s(43) = .62$, $p < .001$, 95% CI [.40, .77], also produced this pattern of results. Taken together, these findings demonstrated that voice-matching performance was largely consistent across the two sessions.

Task-level performance was also investigated using signal detection measures. We found significant associations between T1 and T2 performances for both *d'*, $r_s(43) = .58$, $p < .001$, 95% CI [.35, .75], and *c* values, $r_s(43) = .49$, $p = .001$, 95% CI [.23, .69].

Finally, we investigated whether participants showed any relationship between accuracy and consistency in their performances. As in Experiment 1, accuracy here was calculated as the combined performance (proportion correct) for match and mismatch trials, averaged across the two sessions, whereas inconsistency was calculated as the unsigned difference between the two performances (since larger values reflected greater inconsistency). Across all participants, we found a medium-sized association between accuracy and inconsistency, $r_s(43) = -.32$, $p = .031$, 95% CI [-.56, -.03]. This relationship was also found for match, $r_s(43) = -.47$, $p = .001$, 95% CI [-.67, -.21], but not mismatch trials, $r_s(43) = -.22$, $p = .155$, 95% CI [-.48, .08], when these were analysed separately. In contrast with Experiment 1, no participants' accuracies were greater than or equal to 0.97, and so ceiling accuracies could not account for the relationship found here. Replicating our earlier results with face matching, we found that participants who were more accurate also tended to be more consistent.

3.2.2 Analysis of Ratings

As in Experiment 1, we also analysed the original response ratings. For each participant, separately for each testing session, we calculated their AUC values.

First, we considered whether participant classification performance (AUC) differed across testing sessions. However, a paired-samples *t*-test found no significant difference, $t(44) = 1.63$, $p = .110$, Cohen's $d = 0.24$, 95% CI [-.05, 0.52]. The mean AUC across both sessions was 0.82. Using these values, we also investigated task-level consistency by correlating AUC values for the two sessions. We found a large association between T1 and T2 performance, $r_s(43) = .62$, $p < .001$, 95% CI [.40, .77].

Next, we considered the consistency of participants' responses across the two sessions (i.e., at the trial level). To quantify this, we correlated each participant's T1 and T2 responses, with these values across participants demonstrating substantial within-person agreement, mean $r_s = .54$, 95% CI [.47, .61] (after applying Fisher's *r*-to-*z* transformation and its inverse as necessary).

Finally, we again investigated whether participants showed any relationship between accuracy and consistency in their task-level performances. Here, accuracy was calculated by averaging the AUC values for the two sessions, whereas inconsistency was calculated as the unsigned difference between the two values (with larger values reflecting greater inconsistency). Across all participants, we found a medium-sized association between accuracy and inconsistency, $r_s(43) = -.43$, $p = .004$, 95% CI [-.64, -.16]. Therefore, analysis of the original responses replicated our earlier finding that participants who were more accurate also tended to be more consistent.

3.2.3 Testing Interval and Consistency

Given the range in the number of days (11 to 34) between testing sessions, we considered whether this amount of time was associated with task-level consistency. However, we found no relationship between testing interval and consistency, whether quantified as the unsigned difference between T1 and T2 accuracies using proportion correct, $r_s(43) = .16$, $p = .310$, 95% CI [-.14, .43], or AUC values, $r_s(43) = .09$, $p = .575$, 95% CI [-.21, .37].

3.3 Discussion

In this experiment, we extended our investigation of accuracy and consistency to examine performance on a test of voice matching. To date, there have been no previous studies considering this topic. Our results here are very similar to those of Experiment 1. Again, we find no overall differences in performance across our two sessions – participants were no better or worse at T2. We also find strong support for consistency in performance at the trial level. Finally, we demonstrate a significant association between accuracy and consistency at the participant level, which is medium-sized with both our binary data and when analysing our scale responses. Therefore, as with face matching, we argue that participants who perform more accurately on a test of voice matching also tend to be those who are more consistent.

4 General Discussion

Researchers have demonstrated that matching photographs of unfamiliar faces can be difficult, and people are prone to making mistakes (e.g., Kramer et al., 2018, 2019; Ritchie et al., 2018; White et al., 2014). Of itself, this is an important result since our impressive abilities with familiar faces mean that we are, to some extent, unaware of our limitations with

unfamiliar face matching (Ritchie et al., 2015). Extending this literature, more recent work has begun to explore individual differences in performance since it has become apparent that face matching abilities fall along a continuum (Burton et al., 2010). The aim of the current work was to investigate the nature of these differences, focussing on within-person variability in performance as this consistency is crucial if agencies are to identify suitable candidates for key security roles. A single measure of accuracy fails to provide information regarding whether the individual in question can repeat such a performance, and so we utilised the same testing procedure on two separate occasions. In this way, we were able to explore the relationship between accuracy and consistency: are more accurate individuals also more consistent?

In both our experiments, we found a similar pattern of results. First, there was no overall increase or decrease in performance across the two testing sessions. In other words, there were no practice effects. Simply doing the test again did not help with accuracy, which may be due to the lack of feedback given to participants. Second, a significant degree of consistency was apparent within our individuals, both at the level of the task and trial, suggesting that an underlying, stable ability at the core of their performances resulted in participants scoring higher or lower on the tests, as well as giving the same responses when presented with the same trials. Third, we found a significant association between accuracy and consistency, supporting the notion that these two indices are related for both face and voice matching.

Previous research with face matching has provided mixed evidence regarding the relationship between accuracy and consistency across individuals (Bate et al., 2019; Bindemann et al., 2012). The reasons for this remain unclear, although the main difference between these studies can be found in whether participants were required to complete the same test on two occasions (Bindemann et al., 2012) or different but similar tests (i.e., face

matching blocks focussing on changes in either pose, glasses, or facial hair) once only (Bate et al., 2019). Here, our participants followed the former approach but we found contrasting results to those of Bindemann and colleagues. While Bindemann's analysis suggested a significant, medium-sized association between these two factors (r = .36), removal of their four best observers lowered this correlation so that it was no longer statistically significant (r = .28). Using the same analytical approach, we found comparable associations of -.41 and -.32 respectively (with our negative values simply reflecting the use of inconsistency), with both analyses supporting statistically significant associations. As such, we suggest that the lower sample size (n = 30) in previous work may be the cause for this difference in results. Indeed, we replicated this significant association in our test of voice matching (-.32), as well as in both experiments using AUC measures, again demonstrating that those participants who were more accurate were also more consistent. This conclusion is an important one, given that researchers typically ask individuals to complete a single test of face or voice matching (e.g., Bobak, Dowsett, et al., 2016). Therefore, we argue that the selection of key workers based on single tests is, at least in principle, a justifiable approach, particularly when such recruitment processes may be subject to monetary and time constraints. However, the nature of a specific test used for selection has yet to be determined since the ecological validity of the tests used here remains unknown.

The other notable difference between the studies of Bindemann et al. (2012) and Bate et al. (2019) is that the latter focussed on the performance of police officers who had previously been identified as having proficient face recognition skills. As Bindemann et al. (2012) discuss, the difficulty with quantifying consistency in high performers is that potential ceiling effects necessarily result in higher apparent consistency (i.e., smaller differences across measures). As such, there are clear difficulties with investigating consistency in

potential super-recognisers and other high-accuracy groups, and we suggest the use of more challenging tests of ability in order to address this issue.

For face matching in the current study, we found stronger associations between accuracy and consistency when performance on match and mismatch trials were analysed separately (rather than as overall performance). This result suggests that recruiters can be confident that an individual who scores high on match trials on one occasion, for example, will also perform consistently in the future on this type of trial (Bate et al., 2019). However, for voice matching, we found a stronger association for match trials (in comparison with overall) but a nonsignificant relationship between accuracy and consistency for mismatch trials. We propose that this result is likely due to the lower performance in general on mismatch trials, although further research is required before any conclusions can be drawn.

Strong evidence of consistency in the current set of experiments is made more compelling by the length of interval between testing sessions. Across our participants, a minimum of 11 days passed between T1 and T2 tests. In comparison, Bindemann et al. (2012) asked participants to complete their test of face matching on consecutive days, where the likelihood of recognising specific trials and repeating remembered responses was presumably much higher. Here, we found a correlation of .65 in each of our experiments, demonstrating the large association between overall accuracies at T1 and T2. Indeed, the size of this effect is far greater than those typically found in psychological research (e.g., Gignac & Szodorai, 2016). Interestingly, we also found that interval length was not associated with the level of consistency, i.e., overall performance was not more consistent when the time between testing sessions was lower. Taken together, we argue that individuals are generally consistent in their performance over time, supporting the notion that both face and voice matching represent stable abilities (e.g., Verhallen et al., 2017).

While previous studies have typically utilised 'same'/'different', binary responses when investigating matching tasks (e.g., Bindemann et al., 2012; Burton et al., 2010; Fysh & Bindemann, 2018; Mühl et al., 2018), we asked participants to provide their responses using a 1-5 scale. Although these responses are easily converted to binary-style data for subsequent analysis (O'Toole et al., 2007), we have also explored them in their raw form. The advantage of collecting ratings is that they represent a more fine-grained measure of participants' perceptions. For instance, responding with a 1 followed by a 2 for the same trial over the two sessions would convey the belief that the two faces depicted the same person but represent differing levels of confidence or certainty. Problematically, this change in confidence represents inconsistency that would fail to be detected by simple, binary responses. While a confidence rating alone would detect this perceptual shift, the lack of direction in such a scale would be problematic during analyses since an increase in confidence might accompany more certainty in a "same" response or that the participant is now certain of their change to a "different" response, for example. In addition, our ratings can be used to calculate AUC values, which provide a measure of discrimination for an individual irrespective of their internal threshold. In other words, we can test the quality of the internal value generated (presumably, a judgment regarding the similarity of the two faces along a continuum) rather than the quality of the particular threshold chosen (i.e., which values should result in "same" versus "different" responses?). As mentioned earlier, the threshold for a system can be altered if, for instance, the priority is to avoid 'false alarms'.

By analysing participants' ratings, we found strong evidence of within-person consistency at the trial-level across the two sessions. Simply correlating the two sets of responses demonstrated substantial within-person agreement between T1 and T2 ratings for both face and voice matching (.69 and .54 respectively). While our earlier finding of an association between overall accuracies at T1 and T2 showed that those who score high in one

session are likely to do so again, this result focusses on the trial-by-trial perceptions of the individual. If they are consistent in their perceptions, we should (and do) find a large correlation between their sets of responses.

In Experiment 2, we provide the first evidence that individuals are consistent in their responses when repeating the same voice matching test for the second time. Further, as discussed above, those who are consistent across tests also show higher accuracy. It is both interesting and important that we find the same pattern of results for face and voice matching. With increasing interest in the abilities of super-recognisers and super-matchers (Bate et al., 2019), law enforcement agencies have already begun to recruit and deploy such units in real-world contexts. In recent years, researchers have therefore turned their attention towards how best to identify these individuals (Bate et al., 2018). Although Bindemann and colleagues (2012) argued that it is necessary to measure performance on more than one occasion in order to identify individuals who are consistently accurate, the results of the current work demonstrate that this is not the case. For both faces and voices, we find that selecting super-matchers may be justifiable from performance on a single test (assuming that it demonstrates high construct and ecological validity), which has obvious benefits in an applied setting.

Interestingly, in both experiments, we found consistency in response biases across sessions. That is, when completing the same task at two different timepoints, individual differences in biases were maintained. This result has not previously been reported and suggests that such biases in responding are stable over time. Intuitively, one might predict that biases may simply decrease for all participants at T2, given that they have already been exposed to the stimuli and may have a better sense of what to expect regarding the appearance of match and mismatch trials (although we note that no feedback was given). Somewhat related is the recent finding that providing examples of both trial types (i.e., labelled pairs of match and mismatch faces alongside the current trial) can improve face

matching performance (Gentry & Bindemann, 2019), perhaps allowing participants to better calibrate their judgements. Here, previous exposure to the task provided no increase in performance at T2, with participant differences in response biases remaining stable over time.

In summary, this study examined individual differences in consistency and accuracy for both face and voice matching by asking participants to complete the same test on two separate occasions. The results show that both consistency and accuracy vary across individuals but people are generally consistent in their responses over time, which was evident at both the level of the task and trial. Importantly, we identify a large association between consistency and accuracy, demonstrating that highly consistent individuals also tend to be highly accurate. This information can be used in the identification and recruitment of super-matchers within law enforcement agencies, where such individuals are being used to great effect in a variety of situations on a daily basis.

Data availability statement

The data that support the findings of these experiments are available through the Open Science Framework at https://osf.io/8gtue/?view_only=8ee658e2ad8040adb1977b386c17da9d.

References

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), 22.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019). The consistency of superior face recognition skills in police officers. *Applied Cognitive Psychology, 33*(5), 828-842.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied, 18*(3), 277-291.

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

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

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 339–360.

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*, 207–218.

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

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

Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception, 5*(7), 589-601.

Faul, F., Erdfelder, E., Lang, A. -G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Fysh, M. C. (2018). Individual differences in the detection, matching and memory of faces. *Cognitive Research: Principles and Implications, 3*(1), 20.

Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of Psychology, 109*(2), 219-231.

Gentry, N. W., & Bindemann, M. (2019). Examples Improve Facial Identity Comparison. *Journal of Applied Research in Memory and Cognition, 8*(3), 376-385.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*, 330-337.

Jenkins, R., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2020, January 3). *I remember you: Super-recognisers of faces display superior cross-modal skills with voices*. https://doi.org/10.31234/osf.io/7xdp3

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

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

Kramer, R. S. S., Hardy, S. C., & Ritchie, K. L. (2020). Searching for faces in crowd chokepoint videos. *Applied Cognitive Psychology, 34*(2), 343-356.

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

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

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. London: Chapman & Hall.

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: Implications for criminal investigation and security. *Cognitive Research: Principles and Implications, 3*(1), 26.

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*(3), 576-593.

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), 21.

McConachie, H. R. (1976). Developmental prosopagnosia: A single case report. *Cortex, 12*, 76-82.

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

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

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods, 50*(6), 2184-2192.

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition, 165*, 97-104.

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(9), 1642-1646.

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

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

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

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

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

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), 8.

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

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

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology, 27*(6), 769-777.

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

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.

Tables

Table 1. A summary of performance measures in Experiment 1.

	Session T1	Session T2
Proportion correct	0.86 [0.84, 0.89]	0.87 [0.84, 0.90]
Sensitivity, *d'*	2.45 [2.24, 2.67]	2.61 [2.35, 2.88]
Response bias, *c*	-0.05 [-0.18, 0.07]	-0.05 [-0.17, 0.07]
AUC	0.90 [0.87, 0.92]	0.91 [0.88, 0.94]

Note. 95% confidence intervals are given in square brackets.

Table 2. A summary of performance measures in Experiment 2.

	Session T1	Session T2
Proportion correct	0.79 [0.76, 0.81]	0.77 [0.73, 0.81]
Sensitivity, *d'*	1.80 [1.61, 2.00]	1.78 [1.52, 2.04]
Response bias, *c*	-0.23 [-0.34, -0.11]	-0.31 [-0.46, -0.16]
AUC	0.83 [0.81, 0.86]	0.81 [0.77, 0.85]

Note. 95% confidence intervals are given in square brackets.

Figure captions

Figure 1. Example face pairs from the GFMT. A mismatch pair (top row) and a match pair (bottom row).

Figure 2. A comparison of consistent correct responses (CC), new errors (CE), corrections (EC), and consistent errors (EE) for each trial type in Experiment 1. Error bars represent 95% confidence intervals.

Figure 3. A comparison of consistent correct responses (CC), new errors (CE), corrections (EC), and consistent errors (EE) for each trial type in Experiment 2. Error bars represent 95% confidence intervals.