Fysh, Matthew C. and Bindemann, Markus (2018) *The Kent Face Matching Test.* British Journal of Psychology, 109 . pp. 219-231. ISSN 0007-1269.

The Kent Face Matching Test

Matthew C. Fysh and Markus Bindemann

School of Psychology, University of Kent, Canterbury, UK

Correspondence to:

Matthew C. Fysh, School of Psychology, University of Kent, CT2 7NP, UK.

Email: m.c.fysh@kent.ac.uk

Tel: +44 (0) 1227 823066

Word Count (excluding abstract, references, and figure captions): 4,434

Acknowledgements

We would like to thank Alejandro J. Estudillo and Adam Sandford for assistance with initial image collection for the KUFD.

Abstract

This paper presents the Kent Face Matching Test (KFMT), which comprises 200 same-identity and 20 different-identity pairs of unfamiliar faces. Each face pair consists of a photograph from a student ID-card and a high-quality portrait that was taken at least three months later. The test is designed to complement existing resources for face-matching research, by providing a more ecologically valid stimulus set that captures the natural variability that can arise in a person's appearance over time. Two experiments are presented to demonstrate that the KFMT provides a challenging measure of face matching but correlates with established tests. Experiment 1 compares a short version of this test with the optimised Glasgow Face Matching Test (GFMT). In Experiment 2, a longer version of the KFMT, with infrequent identity mismatches, is correlated with performance on the Cambridge Face Memory Test (CFMT) and the Cambridge Face Perception Test (CFPT). The KFMT is freely available for use in face-matching research.

Keywords: Unfamiliar; face; matching; test

Introduction

In unfamiliar face matching, observers compare pairs of faces to decide whether one person (an identity match) or two different individuals (a mismatch) are depicted. This task is performed routinely at airports, where passport officers verify travellers against their passport photographs. A primary concern in these settings is the detection of impostors (identity mismatches). These are cases in which someone tries to evade detection by travelling under the valid identity documents of another person that is of similar appearance (Stevens, 2011). The true scale of this problem is currently unknown (NCA, 2015). Laboratory studies of face matching have therefore been instrumental in estimating accuracy within such contexts.

A key resource for this research is the Glasgow Face Matching Test (GFMT; Burton, White, & McNeill, 2010). In this test, observers match high-quality, frontal-oriented pairs of faces that are evenly lit and bear neutral expressions. Crucially, identity matches also comprise same-day photographs taken only minutes apart, but with different image-capture devices, to provide optimised conditions to measure best-possible accuracy. Despite such favourable conditions, observers typically record 10-20% errors in this task. This level of performance is already considered problematic for operational settings (Jenkins & Burton, 2008; Robertson, Middleton, & Burton, 2015), but shows also that observers find this task challenging even under ideal conditions.

The GFMT has already featured in over 30 face-matching studies to investigate how performance is impacted by factors such as time pressure (Bindemann, Fysh, Cross, & Watts, 2016), mismatch prevalence (Bindemann, Avetisyan, & Blackwell, 2010), sleep deprivation (Beattie, Walsh, McLaren, Biello, & White, 2016), image quality (Bindemann, Attard, Leach, & Johnston, 2013; Strathie & McNeill, 2016), and performance-related feedback (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014). Moreover, this task has not only been administered to students, but also to non-students (Bobak, Dowsett, & Bate, 2016;

White, Rivolta, Burton, Al-Janabi, & Palermo, 2016), forensic experts (White, Phillips, Hahn, Hill, & O'Toole, 2015), police officers (Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016), and passport officers (White, Kemp, Jenkins, Matheson, & Burton, 2014).

Despite its clear value for psychological research on person identification, the optimised conditions of the GFMT limit the utility of this test under some conditions. For example, one recent study found that face-matching performance can improve when participants collaborate in pairs, but this was only observed with difficult stimuli rather than the optimised faces featured in the GFMT (Dowsett & Burton, 2015). These findings suggest that the optimised conditions provided by the GFMT might obscure some effects that are better identified under more challenging conditions.

Here, we introduce the Kent Face Matching Test (KFMT), which aims to provide such conditions by encapsulating a more applied aspect of face matching. It is currently understood, for example, that face matching is more difficult when to-be-matched stimuli are taken months apart (see, e.g., Megreya, Sandford, & Burton, 2013) or comprise realistic photo-ID images (see, e.g., Bindemann & Sandford, 2011; Kemp, Towell, & Pike, 1997; McCaffery & Burton, 2016). The stimuli of the KFMT are based on such findings, to characterise this dimension of face matching in operational settings, and thus provide a more ecologically valid measure of performance in this task. The KFMT is intended as a complementary resource to be used alongside more optimised measures such as the GFMT, which comprise same-day photographs of identity match pairs, and therefore estimate accuracy only as a best-case scenario.

We first describe the construction of the KFMT and then provide data to compare performance with established tests of face processing. In Experiment 1, we contrast performance between the short versions of the KFMT and GFMT to demonstrate the greater difficulty of our test. This is followed by a second experiment, in which observers completed a longer version of the KFMT, along with two different established tests of face recognition; the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) and the Cambridge Face Perception Test (CFPT; Duchaine, Germine, & Nakayama, 2007).

Test Construction

To assemble the Kent University Face Database (KUFD), we recruited 252 volunteer participants (182 females, 70 males) to have their photograph taken in exchange for a small fee. Each session took place in an evenly lit laboratory, where participants were photographed across various poses and whilst bearing a neutral expression. In the same session, participants were also recorded with a camcorder rotating their heads to look in different directions. Additionally, each participant consented to the use of their Student ID photograph, which was retrieved from the University's online Student Data System. These ID photographs are not constrained by expression, pose, or image-capture device, and therefore represent an important source of variability. The ID photographs were acquired a minimum of three months prior to the laboratory photograph. The mean time interval between acquisition of the laboratory photograph and the ID photograph, across all participants, was approximately 8.8 months (*SD* = 10.5).

With these stimuli, two versions of the KFMT were created. The short version consists of 40 Caucasian identity pairs (20 males, 20 females) from the KUFD. Each pair comprises a high-resolution portrait (Fujifilm FinePix S2980, 14-megapixel) and a student ID photograph. The portrait images were cropped to depict only the target's head and shoulders, rescaled to a size of 283x332 pixels at a resolution of 72-ppi, and were placed on the right hand side of a blank white canvas. The student ID photographs measured 142x192 pixels with a resolution of 72-ppi and positioned to the left of the digital photographs. Thus, each image pair in the KFMT comprises an optimised target photograph taken under controlled conditions, analogous to a passport photograph, but also an ambient photograph in which targets are depicted across a variety of poses, and with different facial expressions. Of the 40 image pairs that feature in the short version of the KFMT, 20 depict the same identity, whilst the remainder depict different individuals. To create these mismatch trials, target images were paired by the experimenters based on their visual similarity with regard to hair colour, face and eyebrow shape. Example match and mismatch pairs are displayed in Figure 1.

The long version of the KFMT comprises 220 face pairs (166 females, 54 males) from the KUFD. Analogous to the short version, identity pairs in this test also comprise a digital portrait, which was cropped to depict only a target's head and shoulders, alongside a student ID photograph. Of these 220 image pairs, 200 depict the same identity, whilst the remainder comprise the 20 identity mismatch pairs that also feature in the short version. In contrast to the short version of the KFMT, which featured only Caucasian faces, some identity pairs in the longer version were also of Asian, Afro-Caribbean, and Middle-Eastern descent. The purpose of this longer test is to further encapsulate the difficulty of face matching conditions in operational contexts such as passport control, by featuring a greater number of trials, and infrequent mismatches.

Experiment 1

This experiment compared performance on the short versions of the KFMT and the GFMT. Normative data show that average performance for the GFMT is around 80-90%, with individual accuracy ranging from near-chance to perfect (see, e.g., Burton et al., 2010). For the KFMT to be a useful resource in face matching research, by providing a more challenging identification test than the GFMT, it is important to establish such a difference in performance.

Method

Participants

Sixty students (40 females, 20 males) from the University of Kent, with a mean age of 20.3 years (*SD* = 3.6), participated in this study in exchange for course credit or a small fee. All participants were British residents and reported normal or corrected-to-normal vision. This study was conducted in accordance with the ethical guidelines of the British Psychological Association.

Stimuli and procedure

The short version of the KFMT was employed for this comparison, as this comprises the same number of identity match trials (20) and mismatch trials (20) as the short version of the GFMT (see Burton et al., 2010). In contrast to the KFMT, one face in each pair of the GFMT consists of a digital photograph image, whilst the other comprises a still-image extracted from high-quality video footage. In each pair, targets are depicted from the front, whilst bearing a neutral expression and under even lighting. All GFMT faces are shown in greyscale, and are presented side-by-side at a width of 350 pixels, with a resolution of 72-ppi.

This experiment was run using *PsychoPy* software (Peirce, 2007). All participants completed the short versions of the KFMT and the GFMT, the order of which was counterbalanced across observers. Each trial was preceded by a 1-second fixation cross, which was then replaced by a stimulus pair. Observers responded using one of two keys on a standard computer keyboard. To measure best-possible accuracy, performance was self-paced in both tasks. No feedback on accuracy was provided during the experiment. As an additional measure to establish the test-retest reliability of the KFMT, 30 participants (16 females, 14 males) from the total sample completed this task twice, with a mean interval of 7.2 days (SD = 0.9) between test sessions.

Results

Response times

For each observer, mean correct response times were calculated for both tests. These are displayed in Figure 2 and suggest that response latencies were comparable between both tasks, with observers on average taking 5.1 and 5.6 seconds to respond on the KFMT and GFMT, respectively. To analyse these data more formally, a 2 (test: KFMT vs. GFMT) x 2 (trial type: match vs. mismatch) within-subjects analysis of variance (ANOVA) was conducted, which did not find an effect of test, $F(1,59) = 0.58$, $p = 0.45$, $\eta_p^2 = 0.01$, or of trial, $F(1,59) = 0.48$, $p = 0.49$, $\eta_p^2 = 0.01$, and these factors did not interact, $F(1,59) = 0.45$, $p = 0.51$, $\eta_p^2 = 0.01$.

Accuracy

Mean percentage accuracy for both tasks was analysed next. This is also displayed in Figure 2 and shows that accuracy in the KFMT was 66% for both match and mismatch trials. By comparison, overall performance in the GFMT was 80%, with slightly higher accuracy on match (82%), compared to mismatch trials (78%). This converges with the baseline level of accuracy for the GFMT in normative studies (e.g., Burton et al., 2010).

To compare performance in these tasks, a 2 (test) x 2 (trial type) within-subjects ANOVA was conducted, which did not reveal an effect of trial type, $F(1,59) = 0.32$, $p = 0.58$, $\eta_p^2 = 0.01$, or an interaction, $F(1,59) = 1.62$, $p = 0.21$, $\eta_p^2 = 0.03$, but showed that accuracy was considerably higher on the GFMT than on the KFMT, $F(1,59) = 104.73$, $p < 0.001$, $\eta_p^2 = 0.64$. This difference is confirmed by an inspection of individual data in Figure 3, which shows that only three of 60 observers performed worse in the GFMT than the KFMT. Despite these differences, overall accuracy correlated for the KFMT and GFMT, $r(58) = 0.45$, $p < 0.001$, which indicates that these tasks measure similar underlying face processes.

Next, the test-retest reliability of the KFMT was analysed. Across sessions 1 and 2, overall accuracy was 66% and 67%, respectively. Overall performance across both test sessions was positively correlated, $r(28) = 0.67$, $p < 0.001$. In addition, a positive relationship was found between sessions for accuracy on match trials, $r(28) = 0.68$, $p < 0.001$, and on mismatch trials, $r(28) = 0.79$, $p < 0.001$. Together, these analyses indicate that the KFMT exhibits high test-retest reliability.

Finally, accuracy was analysed by item to illustrate the range of performance across different face pairs. This is illustrated for match and mismatch stimuli in Figure 4, ordered by item accuracy. These data reiterate that the KFMT is consistently more difficult than the GFMT, and produces a greater range in accuracy across items. In contrast to the GFMT, this range is such that some match pairs are more likely to be classified as identity mismatches, and vice versa.

d-prime and criterion

For completeness, the percentage accuracy data were also converted into signal detection measures of sensitivity (*d'*) and response bias (*criterion*). Paired-sample *t*-tests revealed that *d'* was higher for the GFMT than the KFMT, $t(59) = 10.97$, $p < 0.001$. *Criterion* was comparable for these tests, $t(59) = 1.22$, $p = 0.23$, and was close to zero, both $t\text{s} \leq 1.18$, $p\text{s} \geq 0.24$.

Discussion

This experiment compared performance in a novel test of face matching - the KFMT - with the established GFMT. Accuracy on the KFMT was comparable between match and mismatch trials, and was at 66%. By comparison, overall performance on the GFMT was 80%, with slightly higher accuracy on match (82%) than mismatch (78%) trials. Converging with existing research (e.g., Bindemann, Avetisyan, & Rakow, 2012; Burton et al., 2010; Estudillo & Bindemann, 2014), considerable individual differences emerged in both tests, with accuracy ranging from 40-88% in the KFMT, and 55-100% in the GFMT. However, the difference between tests was robust at an individual level, with only three of 60 participants recording lower accuracy on the KFMT than the GFMT. Performance between tests was positively correlated, which indicates that, despite the increased difficulty of the KFMT, both tests measure similar face processes. In addition, performance on the KFMT also correlated for observers who completed the task twice, with a week's interval between sessions. This suggests that the KFMT reliably measures similar face-matching processes over time.

These data suggest that the KFMT provides a complementary test for the GFMT that could be employed when face-matching accuracy needs to be assessed under more challenging conditions, for example, to mimic more closely applied settings such as passport control. However, in such settings, the number of to-be-matched faces typically exceeds 40 trials, and mismatches occur infrequently. To further understand performance under such conditions, a second experiment was conducted to establish accuracy for the long version of the KFMT. To provide a comparison, this was followed by two other established tests of face processing, the CFMT (Duchaine & Nakayama, 2006) and the CFPT (Duchaine et al., 2007), which measure unfamiliar face recognition and unfamiliar face processing ability, respectively. If the KFMT provides a robust construct, then it should also correlate with these tests.

Experiment 2

In this experiment, observers completed the longer version of the KFMT, comprising 200 match trials and 20 mismatch trials. Current research shows that when matching optimised GFMT faces for a prolonged period, observers develop a response bias to erroneously classify pairs as identity matches (Alenezi & Bindemann, 2013; Alenezi, Bindemann, Fysh, & Johnston, 2015; Bindemann et al., 2016). If the KFMT produces behavioural effects comparable to the GFMT, then such a response bias should also be found here, strengthening the results of Experiment 1.

In addition, observers also completed the CFMT (Duchaine & Nakayama, 2006) and the CFPT (Duchaine et al., 2007) upon completion of the KFMT. In contrast to the face matching task of the KFMT, the CFMT measures recognition memory for newly learned faces, whereas the CFPT requires the ordering of sequences of highly-similar face morphs. However, these three tests are unified on the basis that all focus on the identity processing of unfamiliar faces. The CFMT and CFPT have been used widely and are typically employed to assess impairments in face processing (see, e.g., Bobak, Parris, Gregory, Bennetts, & Bate, 2016; Ulrich, Wilkinson, Ferguson, Smith, & Bindemann, 2016; White et al., 2016), as well as superior recognition ability (Bobak, Hancock, & Bate, 2016; Bobak, Parris, et al., 2016; Russell, Duchaine, & Nakayama, 2009). Thus, the CFMT and CFPT provide suitable tests against which performance in the KFMT can be compared.

Method

Participants

Fifty students (10 males, 40 females) from the University of Kent, with a mean age of 19.5 years (*SD* = 3.0), participated in this study in exchange for course credit. None of these had participated in Experiment 1. All were British residents and reported normal or corrected-to-normal vision.

Stimuli and procedure

KFMT. The long version of the KFMT comprises 220 trials, of which 200 depict the same identity, whilst the remainder are the same 20 mismatch pairs that feature in the short version of this test. These stimuli were evenly divided into four blocks of 55 trials (50 match trials, 5 mismatch trials), which were counterbalanced across observers. Administering the task in this way ensures that mismatch trials were distributed evenly throughout the task, but also allows for the opportunity to observe changes in performance over time. However, to create the impression of one continuous task, no breaks were administered between blocks. As before, this task was run on *PsychoPy* software (Peirce, 2007), with observers responding using one of two keys on a standard computer keyboard. At the beginning of the task, observers were informed that there would be fewer mismatch than match trials. No time pressure or feedback was administered throughout this task, and observers were encouraged to be as accurate as possible.

CFMT. Following the KFMT, observers completed the CFMT (Duchaine & Nakayama, 2006). Stimuli in this task comprise images of six male targets, along with 46 foil identities. All faces are cropped so that features such as hair and facial outline are removed, and depict evenly-lit targets bearing a neutral expression. In the first block of this task, participants study three different orientations of a single target face for three seconds, and are then required to identify the target from a three-face array containing one of the study images and two distractor faces. This is repeated for each target. In the second block, observers study six different but concurrent target faces for 20 seconds, and are then required to identify a given target from a three-face array containing two distractors and a previously-unseen view of a target face. The final block of this task is conceptually similar to Block 2, but with the addition of Gaussian noise over stimuli to further increase the difficulty of this task.

CFPT. Finally, observers completed the CFPT (Duchaine et al., 2007). On each trial, a mid-profile view of a target face is presented, along with six further faces which were created by morphing the target with six individuals by varying amounts. Observers are required to arrange these faces in order of similarity to the target face, with accuracy reflecting the number of deviations from the correct order. This task consisted of 16 trials in total, each of which lasted for a maximum duration of 60 seconds, after which a trial was terminated and the next was initiated. Additionally, half of these trials depicted upright faces, which were randomly intermixed with the remaining eight trials, in which faces were presented upside-down.

Results

KFMT

Response times

Mean correct response times were analysed first and are displayed in Figure 5. These reflect that match and mismatch response times were comparable throughout the task. To analyse these data formally, a 2 (trial type: match vs. mismatch) x 4 (block: 1, 2, 3, 4) within-subjects ANOVA was conducted, which revealed an effect of trial type, $F(1,49) = 6.76$, $p < 0.05$, $\eta_p^2 = 0.12$, due to faster responses on match trials than on mismatch trials. In addition, an effect of block was found, $F(3,147) = 4.03$, $p < 0.01$, $\eta_p^2 = 0.08$. However, none of the pairwise comparisons between blocks were significant following Bonferroni adjustment, all ps ≥ 0.08. These factors did not interact, $F(3,147) = 0.45$, $p = 0.71$, $\eta_p^2 = 0.01$.

Accuracy

Average accuracy for match and mismatch trials across blocks was 78% and 64%, respectively. However, the data depicted in Figure 5 reflect that over Blocks 1 through 4, performance on mismatch trials deteriorated from 74% to 57%, whereas accuracy on match trials increased from 71% to 82%. A 2 (trial type) x 4 (block) within-subjects ANOVA was conducted to investigate this variation in performance across blocks. This did not reveal a main effect of block, $F(3,147) = 1.19$, $p = 0.32$, $\eta_p^2 = 0.02$, but of trial type, $F(1,49) = 7.90$, $p < 0.01$, $\eta_p^2 = 0.14$, and a significant interaction, $F(3,147) = 10.64$, $p < 0.001$, $\eta_p^2 = 0.18$.

Analysis of simple main effects revealed that this was due to a deterioration in accuracy on mismatch trials, $F(3,47) = 4.00$, $p < 0.05$, $\eta_p^2 = 0.20$, which was lower in Blocks 3 and 4 compared to Block 1, both *p*s < 0.05, but was comparable between all other blocks, all *p*s ≥ 0.08. The improvement in performance on match trials across blocks was also significant, $F(3,47) = 11.19$, $p < 0.001$, $\eta_p^2 = 0.42$, with higher accuracy in the final block compared to all other blocks, all *p*s < 0.05, as well as in Blocks 2 and 3 compared to Block 1, both *p*s < 0.001. However, performance was similar between Blocks 2 and 3, $p = 1.00$. In addition, performance on match and mismatch trials was comparable in the first block, $F(1,49) = 0.34$, $p = 0.56$, $\eta_p^2 = 0.01$, but accuracy on match trials was superior in the second, $F(1,49) = 6.15$, $p < 0.05$, $\eta_p^2 = 0.11$, third, $F(1,49) = 8.13$, $p < 0.01$, $\eta_p^2 = 0.14$, and final block, $F(1,49) = 14.72$, $p < 0.001$, $\eta_p^2 = 0.23$.

d-prime and criterion

Percentage accuracy scores were also converted into *d'* and *criterion*. A one-factor within-subjects ANOVA did not reveal an effect of block for *d'*, $F(3,147) = 0.28$, $p = 0.84$, $\eta_p^2 = 0.01$, but for *criterion*, $F(3,147) = 12.45$, $p < 0.001$, $\eta_p^2 = 0.20$. Pairwise comparisons revealed that this reflects a greater tendency to classify face pairs as identity matches in

Blocks 2, 3, and 4, compared to Block 1, all ps < 0.01. To confirm this bias, one-sample t-tests were conducted to compare *criterion* to zero in each block. This revealed that *criterion* was comparable to zero in the first block, $t(49) = 0.76$, $p = 0.45$, but was reliably below zero in the second, $t(49) = 2.57$, $p < 0.05$, third, $t(49) = 2.98$, $p < 0.01$, and final block, $t(49) = 3.91$, $p < 0.001$.

CFMT and CFPT

Accuracy

Overall accuracy on the CFMT was at 76%, which is comparable to the average score of 80% in its normative tests (see, Duchaine & Nakayama, 2006). Average performance was at ceiling in the first block of this test, with 98% correct identifications, but deteriorated to 73% and 62% in Blocks 2 and 3, respectively.

On the CFPT, the average number of deviations (errors) from the correct order across all trials was 51.4. Performance was considerably better on upright than on inverted trials, with 35.5 versus 67.2 deviations, $t(49) = 13.75$, $p < 0.001$. The number of errors in the CFPT correlated negatively with accuracy on the CFMT, $r(48) = -0.41$, $p < 0.01$, reflecting that face memory is positively associated with face perception ability.

Correlations with the KFMT

To explore whether variation in performance on the KFMT was reflective of general ability in face memory and face processing, a correlation analysis was performed with the CFMT and CFPT (see Figure 6). This revealed a positive relationship between the KFMT and CFMT, $r(48) = 0.29$, $p < 0.05$, and a negative relationship between accuracy on the KFMT and the number of errors in the CFPT, $r(48) = -0.34$, $p < 0.05$.

Discussion

In this experiment, observers completed a longer version of the KFMT, as well as the CFMT and the CFPT. Overall performance in the KFMT was 70%, with 78% and 64% accuracy for match and mismatch trials, respectively. This is slightly higher than on the short version of the KFMT, in which overall accuracy was 66%. However, in the long version of this task, accuracy on mismatch trials deteriorated substantially across blocks, from 74% to 57%. Conversely, performance on match trials improved, from 72% to 82%. This pattern is reflected by a shift in *criterion*, which indicates that a response bias to classify an increasing number of faces as identity matches emerged over time. Such a bias has also been found in recent work using the optimised stimuli of the GFMT, but with initial accuracy levels that exceed 80% (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann et al., 2016). The data from Experiment 2 therefore converge with Experiment 1 to indicate that the KFMT provides a more challenging test for face matching than the GFMT, but preserves the behavioural characteristics of this test. In addition, accuracy in the KFMT also correlated with the CFMT and CFPT (see Duchaine & Nakayama, 2006; Duchaine et al., 2007). This demonstrates further that unfamiliar face matching performance on the KFMT utilises mechanisms similar to those employed for unfamiliar face memory on the CFMT and unfamiliar face perception in the CFPT.

General Discussion

This study presents the KFMT as a new test of face matching and examined its characteristics across two experiments. Performance on the KFMT correlated with the GFMT in Experiment 1, and also followed the accuracy profile that is found over longer experiments with this test in Experiment 2 (see, Alenezi & Bindemann, 2013; Alenezi et al., 2015). This indicates that both tests measure similar underlying processes. However, face-matching accuracy was substantially lower on the KFMT than on the GFMT, by 14%, and this effect was robust on both an individual level and by item. In addition, performance on the short version of the KFMT correlated for observers who completed this test one week apart. This demonstrates that this task measures the same processes between separate testing sessions with high reliability. Finally, Experiment 2 showed that performance on the KFMT was associated with the CFMT and CFPT, which also measure aspects of unfamiliar face-identity processing.

Taken together, these results indicate that the KFMT is a psychometrically-stable test of unfamiliar-face matching, but the variability in the face photographs of its stimulus pairs provides a more challenging identification test than the established GFMT, which is based on optimised stimuli for person identification. We note that these conclusions are based on samples here that feature an unequal sex ratio. However, sex differences exert only a numerical effect of around 5% on face-matching performance (see Megreya & Bindemann, 2015), which is small compared to the very broad individual differences in face-matching accuracy between observers of the same sex (see, e.g., Burton et al., 2010; Megreya & Bindemann, 2013).

The aim of this paper is to facilitate further research with the KFMT to understand face-matching performance in the context of passport control. We suggest that this makes the KFMT a valuable research resource to investigate factors that cannot be explored fully with the optimised identification conditions that are provided by the GFMT (see, e.g., Bindemann et al., 2016; Dowsett & Burton, 2015). The KFMT is freely available for face-matching research (via www.kent.ac.uk/psychology/people/bindemannm/index.html).

References

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

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, e1184. doi:10.7717/peerj.1184

Beattie, L., Walsh, D., McLaren, J., Biello, S. M., & White, D. (2016). Perceptual impairment in face identification with poor sleep. *Open Science*, *3*, 160321. doi:10.1098/rsos.160321

Bindemann, M., Avetisyan, M., & Blackwell, K.-A. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*, 378-386. doi:10.1037/a0021893

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

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied, 18,* 277-291. doi: 10.1037/a0029635

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar face matching. *Applied Cognitive Psychology, 27*, 707-717. doi:10.1002/acp.2970

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching faces against the clock. *i-Perception*, *7*, 2041669516672219. doi: 10.1177/2041669516672219

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*, e0148148. doi:10.1371/journal.pone.0148148

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*, 81-91. doi: 10.1002/acp.3170

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2016). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *The Quarterly Journal of Experimental Psychology*, *0*, 1-17. doi:10.1080/17470218.2016.1161059

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*, 286-291. doi:10.3758/BRM.42.1.286

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*, 433-445. doi:10.1111/bjop.12103

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*, 576-585. doi:10.1016/j.neuropsychologia.2005.07.001

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*, 419-430. doi:10.1080/02643290701380491

Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, *5*, 589-601. doi:10.1068/i0669

Jenkins, R., & Burton, A. M. (2008). Limitations in facial identification: The evidence. *Justice of the Peace, 172*, 4-6.

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

McCaffery, J. M., & Burton, A. M. (2016). Passport checks: Interactions between matching faces and biographical details. *Applied Cognitive Psychology*. doi:10.1002/acp.3281

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology, 25*, 30-37. doi:10.1080/20445911.2012.739153

Megreya, A. M., & Bindemann, M. (2015). Developmental improvement and age-related decline in unfamiliar face matching. *Perception, 44*, 5-22. doi:10.1068/p7825

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*, 700-706. doi:10.1002/acp.2965

National Crime Agency (NCA). (2015). *National strategic assessment of serious and organised crime 2015.* Retrieved March 2, 2016, from http://www.nationalcrimeagency.gov.uk/publications/560-national-strategic-assessment-of-serious-and-organised-crime-2015/file

Peirce, J. W. (2007). PsychoPy – Psychophysics software in python. *Journal of Neuroscience Methods, 162,* 8-13. doi:10.1016/j.jneumeth.2006.11.017

Robertson, D., Middleton, R., & Burton, A. M. (2015). From policing to passport control. *Keesing Journal of Documents & Identity, February*, 3-8.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face Recognition by Metropolitan Police Super-Recognisers. *PloS one*, *11*, e0150036. doi:10.1371/journal.pone.0150036

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

Stevens, C. (2011). *Facing up to impostor fraud*. Retrieved May 26, 2014, from http://www.icao.int/publications/journalsreports/2011/ICAO%20MRTD%20Report%20Vol.6%20No.3,%202011.pdf

Strathie, A., & McNeill, A. (2016). Facial wipes don't wash: Facial image comparison by video superimposition reduces the accuracy of face matching decisions. *Applied Cognitive Psychology, 30,* 504-513. doi:10.1002/acp.3218

Ulrich, P. I., Wilkinson, D. T., Ferguson, H. J., Smith, L. J., Bindemann, M., Johnston, R. A., & Schmalzl, L. (2016). Perceptual and memorial contributions to developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology*, 1-18. doi:10.1080/17470218.2016.1177101

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

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

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. In *Proc. R. Soc. B* (Vol. 282, No. 1814, p. 20151292). The Royal Society. doi:10.1098/rspb.2015.1292

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2016). Face matching impairment in developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology*, 1-11. doi:10.1080/17470218.2016.1173076

FIGURE 1. Example match (top row) and mismatch (bottom row) pairs from the KFMT.

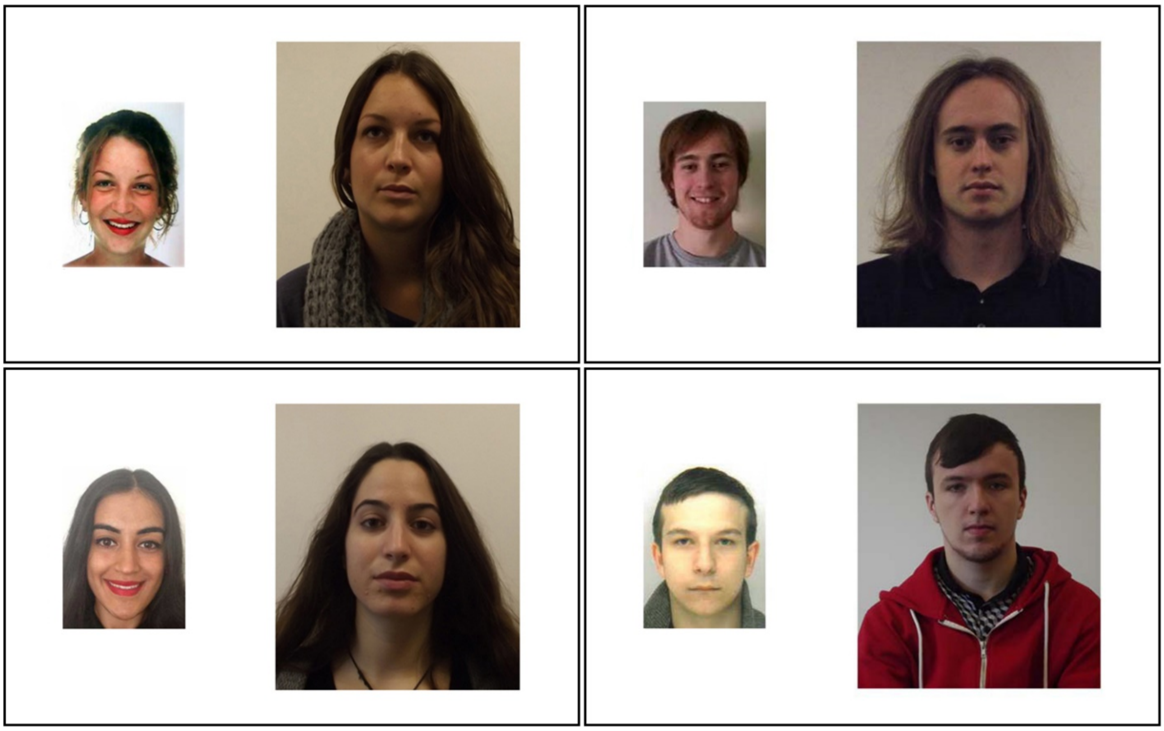

FIGURE 2. Mean correct response times (top) and percentage accuracy scores (bottom) for match and mismatch trials in the short versions of the KFMT and GFMT.

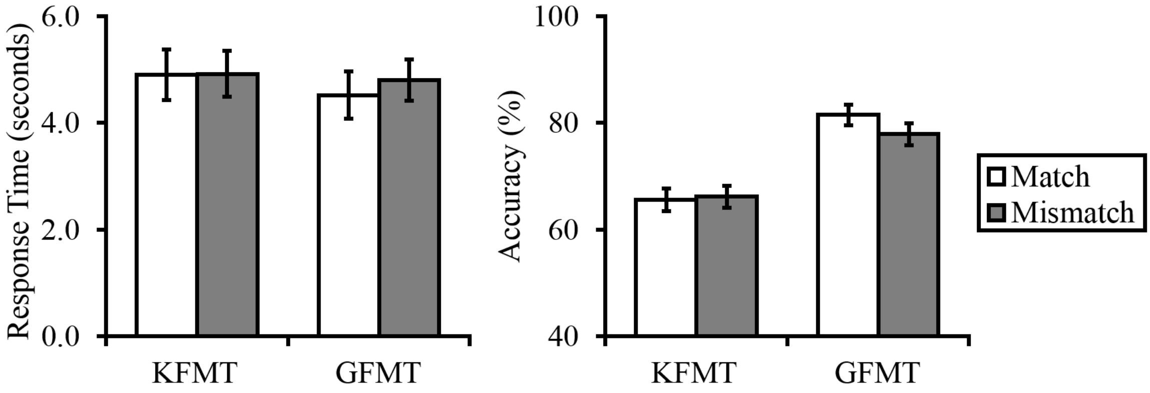

FIGURE 3. Individual data, based on overall accuracy, for the KFMT and GFMT, ordered from least to most accurate observer.

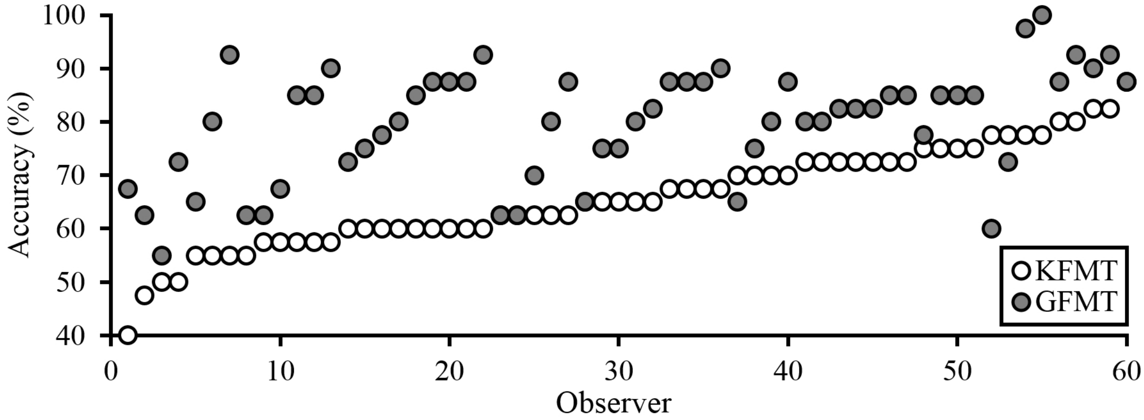

FIGURE 4. Percentage accuracy for individual items on the KFMT and the GFMT, ordered from least to most accurate.

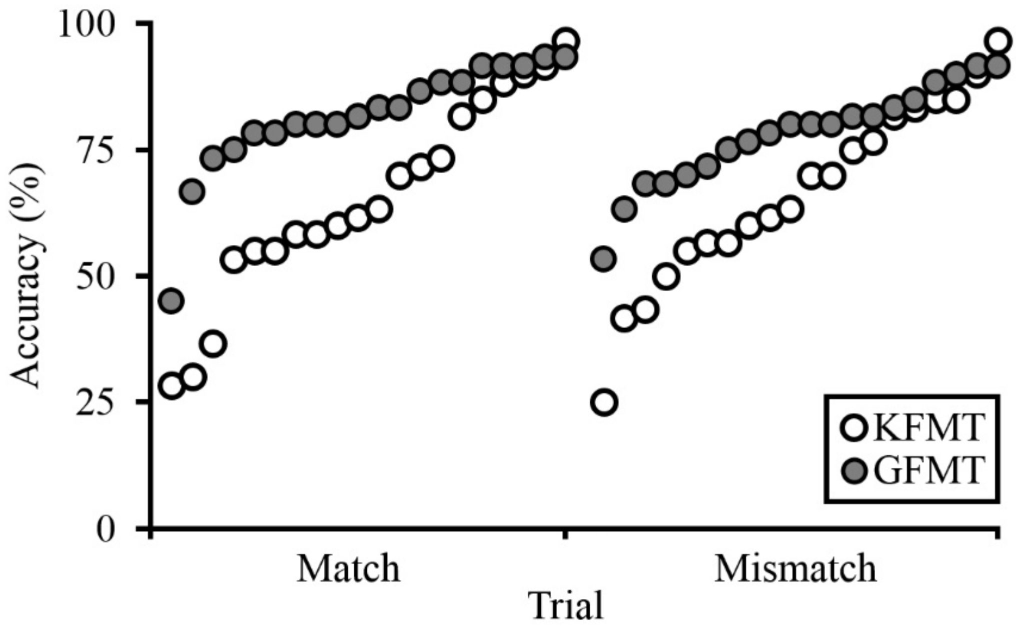

FIGURE 5. Percentage accuracy, mean correct response times, *d'*, and *criterion* on the long version of the KFMT. Open markers denote match trials, and grey markers denote mismatch trials. Error bars denote the standard error of the mean.

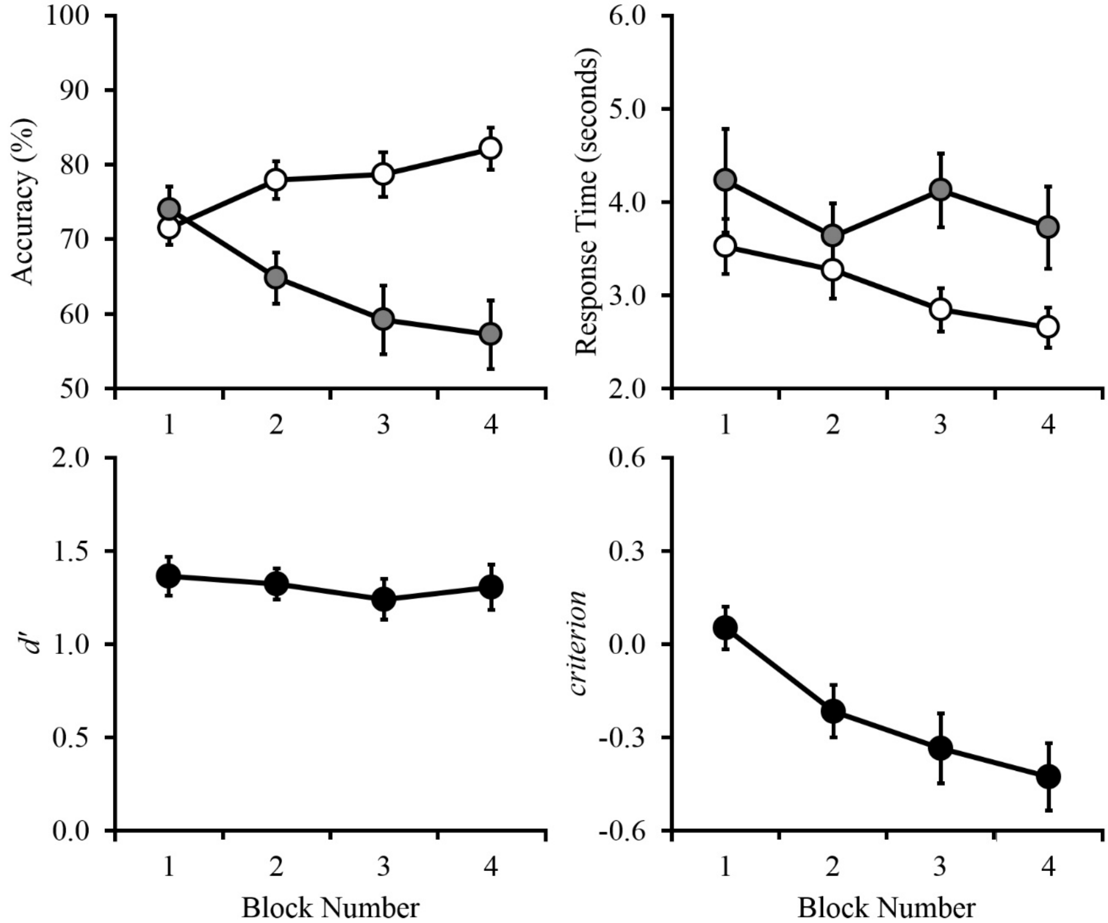

FIGURE 6. Scatter plots for overall performance on the KFMT versus the CFMT and the CFPT.