FACE IDENTIFICATION IN THE LABORATORY AND IN VIRTUAL WORLDS

Markus Bindemann[1], Matthew C. Fysh[1], Iliyana V. Trifonova[2], John Allen[1],

Cade McCall[2], & A. Mike Burton[2]

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

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

Correspondence to:

Professor Markus Bindemann

School of Psychology, University of Kent, CT2 7NP, UK.

Email: m.bindemann@kent.ac.uk / Tel: +44 (0) 1227 823087

Word count (excluding abstract, method, results, figure captions and references): 3,476

Abstract

Investigations into human cognition typically control variables tightly in the laboratory or relinquish systematic control in field studies. Virtual reality (VR) can provide an intermediate approach, by facilitating research with complex but controlled environments. However, understanding of the correspondence between VR and laboratory paradigms is still limited. This study addresses this issue by comparing established laboratory tests of face identification with passport control at a VR airport. We show that test characteristics transcend comparison of the laboratory tests and VR, and demonstrate consistent correlations between these tasks. However, person identification in VR was also marked by bias to accept mismatching identities. These findings support correspondence between laboratory tests of face perception and VR, but also highlight the importance of understanding human behaviour under more complex conditions. This problem arises in many areas of psychology and our study shows that VR offers a solution, by providing complex but controlled environments.

Keywords: virtual reality; person identification; face matching; passport control; airport

General Audience Summary

Psychological experiments into human cognition either tend to study behaviour in the laboratory, where the conditions under which research is conducted are simplistic but tightly controlled, or relinquish such control in field studies, where behaviour is examined in natural environments in which additional factors can be at play. Both approaches have some disadvantages that the development of Virtual Reality (VR) can bridge, by facilitating behavioural research in environments that are both complex and controlled. However, how research in VR corresponds to traditional experimental approaches is still unknown. This study investigates this issue by comparing established laboratory tests of face identification with person identification at a VR airport, in which participants take on the role of passport control officers. We demonstrate that person identification in laboratory tests is linked to the same behaviour in VR. However, we also find that person identification in VR is marked by a tendency to incorrectly accept travellers who bear the identity documents of another person. These findings demonstrate the importance of understanding human behaviour under conditions that more closely mimic real life and show that VR can facilitate such research.

Introduction

In Psychology, laboratory investigations into human cognition typically control variables tightly to isolate processes of interest. However, this control must be balanced against the real-world behaviours that such experiments are seeking to address. This issue arises in many areas of psychology, and in this paper, we consider the problem of face perception. If laboratory tasks do not preserve important characteristics of the environment and social contexts within which behaviour occurs, then we risk developing theories from experimental data which fail to adequately explain real-world cognition. This balance between exercising tight experimental control and capturing complex behaviour is difficult to achieve, and so a dichotomy has emerged in the study of cognition. On one side, laboratory experiments typically provide impoverished contexts, in which stimuli are presented in highly simplified displays. In these experiments, the real-world contexts within which behaviours occur are not considered essential in understanding a process. The alternative approach to these laboratory experiments are field studies. These acknowledge the importance of context for understanding behaviour, but also relinquish systematic control over the variables that are at play. This poses a conundrum for researchers. How is it possible to conduct behaviourally-relevant research under conditions that also provide the necessary experimental control to isolate variables of interest?

A potential answer to this problem comes from the development of affordable Virtual Reality (VR) systems, which enable researchers to immerse participants in environments that are increasingly complex and realistic, but which also preserve the controlled nature of laboratory experiments (Loomis et al., 1999; McCall & Blascovich, 2009; McCall et al., 2016; Wilson & Soranzo, 2015). This provides a means of studying behaviour in scenarios that are impossible to simulate effectively in the laboratory, but which are also difficult to access or control in the field. However, the implementation of VR also requires some

fundamental changes to the way in which experiments are typically conducted, and investigations into the correspondence between VR and laboratory paradigms are still limited. It is therefore important to establish whether some experimental approaches transfer naturally to implementation in VR. In particular, are key aspects of the natural world preserved in the transition to VR, such that experiments within this artificial environment provide useful and generalisable results?

In this study, we investigate this approach for studying *face perception*, a popular topic in modern psychology (for reviews, see Bruce & Young, 1998; Hole & Bourne, 2010; Bindemann & Megreya, 2017; Rhodes et al., 2011). For research in VR, faces must be rendered onto human avatars (e.g., Bailenson et al., 2003, 2008; Bülthoff et al., 2019). Here we seek to establish the correspondence of the perception of such avatars in a complex VR environment with laboratory-based methods that show photographs of real faces in simplified displays. This is an important step to establishing the behavioural relevance of VR to psychological experimentation with faces.

Specifically, we focus on the task of unfamiliar face matching, which requires observers to determine whether two faces depict the same person or different people. In the laboratory, this task is performed using pairs of face photographs, which are typically presented in isolation. This simple task is considered an analogue to the identification of travellers at airports and borders, and has been studied extensively in recent years (see Bindemann, 2021). Much of this research has attempted to understand applied aspects of this task. This work has shown, for example, that face matching is difficult for lay persons and passport control professionals alike (Towler et al., 2019; White et al., 2014; White, Dunn, et al., 2015; Wirth & Carbon, 2017), but is also marked by large individual differences in novices (Burton et al., 2010; Fysh & Bindemann, 2018) and practitioners (Phillips et al., 2018; White et al., 2014; see Lander et al., 2018, for a review). And while some professionals

excel at face matching (Towler et al., 2017; White, Dunn, et al., 2015), screening for such professionals using conventional laboratory methods is difficult (Bate et al., 2018; Fysh et al., 2020). This difficulty is compounded further by studies showing that real-world factors, such as passenger volume and time pressure, influence face-matching accuracy in the laboratory (Bindemann et al., 2016; Fysh & Bindemann, 2017; Wirth & Carbon, 2017; for a review of factors, see Fysh, 2021). However, while researchers can use laboratory data to estimate how these factors influence performance in practice, unfamiliar face matching has not been studied in the critical real-world context of passport control at airports, due to the security-sensitive nature of these environments.

VR provides a solution to this problem by circumventing the access issues that exist in the real world. We have recently developed a VR airport-based method to study face identification at passport control (Tummon et al., 2019), and demonstrated the potential of this approach to investigate contextual factors in ways that extend laboratory research in this domain (Tummon et al., 2020). Alongside this innovation, we have developed methodology to construct avatars with photo-realistic faces for psychological experimentation in VR (Fysh et al., in press). Here, we combine these approaches to study identification of realistic person avatars at the passport control checkpoint of a VR airport. For this purpose, an avatar was paired with a face photograph at passport control and observers determined whether these pairings depict the same person.

We compare identification performance in this airport task with two established laboratory tests of face matching. The first of these is the Glasgow Face Matching Test (GFMT), which examines identity matching under optimised viewing conditions, by utilising highly-controlled face images of the same person that were captured only a few moments apart (Burton et al., 2010). This test construction is similar to identification in our VR airport, which is also based on the comparison of avatars' faces with a same-day face photograph.

This is contrasted with the Kent Face Matching Test (KFMT), in which face pairs consist of a controlled face portrait and an uncontrolled image that were captured several months apart, thus providing a more challenging identification task (Fysh & Bindemann, 2018). To investigate the link between these laboratory tests and person identification in virtual worlds, we compared group-level accuracy on the GFMT and KFMT with that of avatars to determine whether test characteristics transcend the comparison of controlled laboratory tests and VR. This was complemented by examining whether individual performance was consistent across the laboratory tests and person identification in VR, by correlating accuracy across these tests.

Experiment 1

We begin by comparing performance on the GFMT and KFMT with the identification of avatar faces when these are presented without the airport context. In this way, we seek to establish whether the graphic-based avatar faces support face matching in a way that corresponds to well-established photo-to-photo comparison methods. To do this, an image of each avatar face was paired with a face photograph, and observers were asked to identify these pairings as identity matches or mismatches. We then compared average identification accuracy for the GFMT, KFMT and avatars, and examined whether these tests are associated. This provides an important first step towards understanding correspondence of these tasks before the avatars are presented in an airport context in Experiments 2 and 3.

Method

This experiment was preregistered on the Open Science Framework (OSF) (https://osf.io/jcu74/?view_only=).

Participants

Ninety-six participants were initially recruited for this experiment. However, five of these were excluded because they failed more than 25% of attention check trials, and one other participant was excluded due to giving the same response to all experimental trials. Our final sample therefore consisted of 90 participants (62 females, 28 males) with a mean age of 30.8 years (*SD* = 10.0) who were recruited for this study in exchange for a small fee using Prolific Academic. All participants resided in the United Kingdom at the time of recruitment. Our sample size was guided by studies of a similar nature (e.g., Fysh & Bindemann, 2018), however we increased our target sample size by 50% to tolerate additional noise that can arise from online experimentation (e.g., Ramon, 2021).

Stimuli

Each participant completed three tasks, comprising of the GFMT, KFMT and an avatar face matching test, which are described below. The presentation of these tests was blocked and the order counterbalanced across participants. Example stimuli for all tests are shown in Figure 1.

Glasgow Face Matching Test: This experiment employed 20 identity match and 20 mismatch trials from the short version of the GFMT. These consist of pairs of face images (all Caucasian) recorded from a frontal view while displaying a neutral expression. Each image in a face pair was taken with different cameras and, in the case of identity matches, approximately 15 minutes apart. Each face image was cropped to show the head only, converted to greyscale, and sized to 350 pixels in width at a resolution of 72 ppi (for detailed information on the GFMT, see Burton et al., 2010). Half of the stimuli depicted male faces, and the remaining half depicted female faces (10 per match/mismatch condition). Because

data collection was conducted online, four additional trials (2 matches, 2 mismatches) were also included, in which face pairs were presented upside down, as an attention check.

Kent Face Matching Test: The KFMT face pairs consist of an image from a student ID card, sized at 142 pixels in width, and a portrait photo, sized at 283 pixels width at a resolution of 72 ppi. The student ID photos were taken at least three months prior to the face portraits and were not constrained by pose, facial expression, or image-capture device. The portrait photos depict the target's head and shoulders from a frontal view whilst bearing a neutral facial expression and were captured with a high-quality digital camera. In this experiment, 20 identity match and 20 mismatch trials from the short version of the KFMT were employed, with 10 males and 10 females per match/mismatch condition.[1] All faces were Caucasian (for more information, see Fysh & Bindemann, 2018). As with the GFMT, four additional trials were included (2 matches, 2 mismatches), in which faces were presented upside down, as an attention check.

Avatar Face Matching Test: Forty avatars paired with high-quality digital photographs served as stimuli for this experiment (20 matches and 20 mismatches). Each avatar was constructed using a 3D scan of a real person's head that was acquired using a state-of-the-art 3D scanner, which was subsequently rigged onto a pre-made body and animated for movement in VR (see Fysh et al., in press). For each avatar, a high-resolution screenshot was acquired, which was subsequently paired alongside a high-quality digital photograph of the scanned subject's real-life counterpart (i.e., a match trial) or a different person who was matched for gender and approximate age and broadly similar in appearance (i.e., a mismatch trial). Half of the face pairings depicted male and half depicted female faces. These images were size to 200 pixels width at a resolution of 72 ppi. None of the identities

[1] Although long versions of the GFMT and KFMT are available, with 168 and 220 face pairs, we employed the short versions as these are matched for length and match-mismatch ratio.

were repeated across face pairings, resulting in 40 unique stimulus arrays. Once again, four additional trials (2 matches, 2 mismatches) were employed as an attention check, in which the faces were presented upside down, for which observers were required to press the spacebar.

Procedure

The three tests were programmed with *Gorilla Experiment Builder* (Anwyl-Irvine et al., 2020), which was also used to collect the data online. All three face tests followed the same procedure. A trial began with a 1-second fixation cross, followed by a face pair, which remained on display until a response was registered. Participants were asked to categorize these face pairs as identity matches or mismatches via two button presses ('S' and 'D') on a computer keyboard. In addition, a small proportion of face pairs were presented upside-down as attention checks, requiring responses via a third key (space bar) irrespective of the match/mismatch nature of the face pair. In this manner, participants were first presented with a short practice block of 10 trials (4 matches, 4 mismatches, 2 attention check trials), which was based on cartoon faces. This was followed by 44 experimental trials (20 matches, 20 mismatches, 4 attention checks) for each test (GFMT, KFMT, Avatars). Presentation of stimuli was randomised for each participant. Performance was self-paced and participants were instructed to respond as accurately as possible. The presentation software adjusted presentation size of all face pairs depending on the screen size of the participant. For example, on a 17″ monitor (1920 x 1080 pixels / 38.3 x 21.5 cm), the faces in the GFMT measured approximately 40 (w) x 50 (h) mm, the ID photo in the KFMT measured 28 (w) x 39 (h) mm and the face portrait 56 (w) x 66 (h) mm, and all faces in the Avatar test measured 40 (w) x 50 (h) mm. Permitted devices to complete the experiment were limited to laptop and desktop computers.

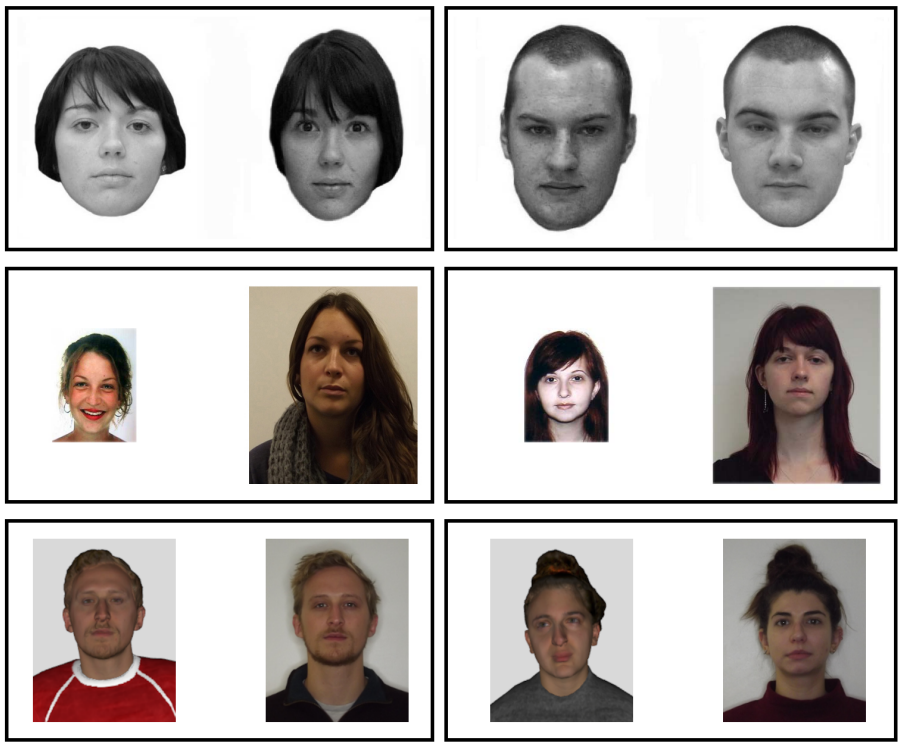

FIGURE 1. Example stimuli for the GFMT (top), KFMT (middle) and Avatar test (bottom) in Experiment 1, showing identity matches (left column) and mismatches (right column).

Results

Accuracy

To compare group-level performance, mean accuracy was calculated for match and mismatch trials for the three tasks. These data are illustrated in Figure 2. A 3 (test: GFMT, KFMT, Avatars) x 2 (trial type: matches, mismatches) within-subjects ANOVA of these data did not show a main effect of trial type, $F(1,89) = 0.31$, $p = .58$, $\eta_p^2 = .00$, but a main effect of test, $F(2,178) = 128.89$, $p < .001$, $\eta_p^2 = .59$, and an interaction between factors, $F(2,178) = 7.04$, $p < .01$, $\eta_p^2 = .07$.[2]

[2] RTs for correct responses are reported for completeness. A 3 (task) x 2 (trial type) ANOVA of RTs revealed an interaction, $F(2,176) = 4.21$, $p < .05$, $\eta_p^2 = .05$, due to a simple main effect of task for matches, $F(2,87) = 14.78$, $p < .001$, $\eta_p^2 = .25$, but not mismatches, $F(2,87) = 1.54$, $p = .22$, $\eta_p^2 = .03$. Matches were classified more quickly than mismatches on the Avatar test (2510 vs. 2906), $F(1,88) = 6.80$, $p < .05$, $\eta_p^2 = .07$, but not the GFMT (2869 vs. 2968), $F(1,88) = 0.73$, $p = .40$, $\eta_p^2 = .01$, or KFMT (3203 vs. 3141), $F(1,88) = .41$, $p = .53$, $\eta_p^2 = .01$. In

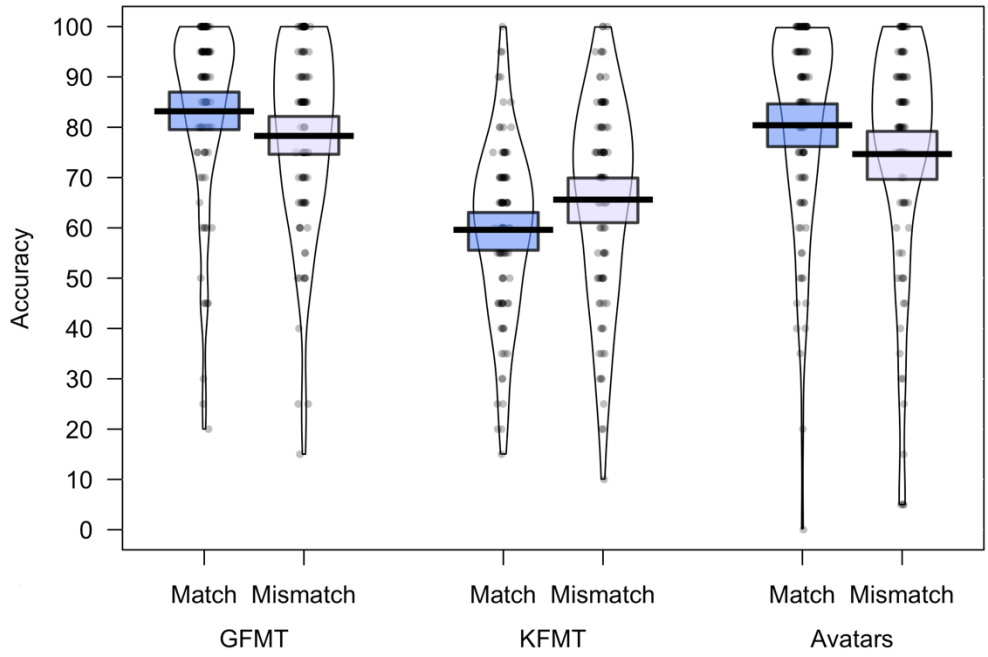

FIGURE 2. Accuracy (%) for match and mismatch trials across the three tasks employed in Experiment 1. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

Analysis of simple main effects revealed a difference in accuracy across tests for matches, $F(2,88) = 98.58$, $p < .001$, $\eta_p^2 = .69$, and mismatches, $F(2,88) = 24.08$, $p < .001$, $\eta_p^2 = .35$. For both matches and mismatches, Tukey HSD test showed that accuracy was higher for the GFMT and the Avatars than for the KFMT, all ps $< .001$. In contrast, performance was comparable for the GFMT and Avatars on match, $p = .78$, and mismatch trials, $p = .53$. Finally, accuracy for matches and mismatches did not differ reliably for the GFMT, $F(1,89) = 3.22$, $p = .08$, $\eta_p^2 = .04$, the KFMT, $F(1,89) = 3.08$, $p = .08$, $\eta_p^2 = .04$, or the Avatars, $F(1,89) = 2.10$, $p = .15$, $\eta_p^2 = .02$. Overall, these data therefore show that group-level performance was similar for the GFMT and the Avatars, and better for these two tests than the KFMT.

addition, match RTs were faster on the Avatar test than the KFMT, $p < .001$, but more comparable between the Avatar test and GFMT, $p = .08$, and between the KFMT and GFMT, $p = .14$ (Tukey HSD).

This is consistent with the design of the KFMT to be harder than GFMT, and shows good levels of performance for Avatars, consistent with the easier of the two photo-to-photo tests.

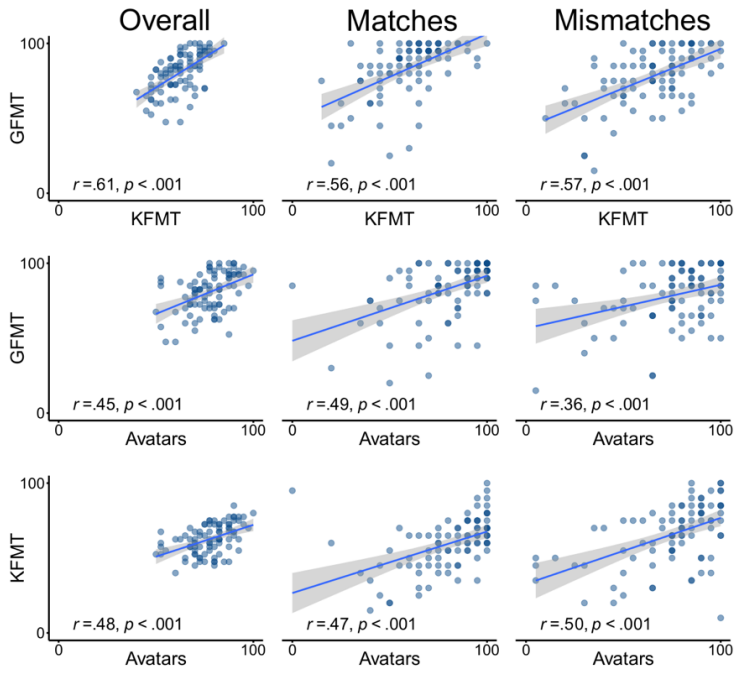

FIGURE 3. Accuracy (%) correlations between the GFMT, KFMT, and the Avatar matching task for overall, match and mismatch accuracy in Experiment 1.

To examine associations in individual identification performance across the GFMT, KFMT and Avatars, pairwise Pearson's correlations were performed between all tests. These data are illustrated in Figure 3 and show positive correlations between the GFMT and KFMT in overall, match and mismatch accuracy, all *rs* ≥ .56, *ps* < .001. Similarly, correlations were observed between the GFMT and the Avatar test for the three accuracy measures, all *rs* ≥ .36, *ps* < .001, as well as the KFMT and the Avatar test, all *rs* ≥ .47, *ps* < .001[3,4]. Taken together,

[3] These correlational relationships did not meaningfully change following the removal of an outlying participant, who scored 100% match accuracy on the KFMT and 0% match accuracy on the Avatar test (see Figure 3), with all *r*s ≥ .44, *p*s < .001 for correlations of overall accuracy between tests, all *r*s ≥ .54, *p*s < .001 for correlations of match accuracy, and all *r*s ≥ .39, *p*s < .001 for correlations of mismatch accuracy.

[4] To guard against false positive correlations, the Benjamini-Hochberg procedure was also applied to check each *p*-value against a critical threshold that was determined by the number of comparisons. All relationships

these data therefore show moderate correlations in individual accuracy across all test comparisons and all measures, indicating that identification performance is associated across the three tasks.

d' and criterion

Accuracy was also converted into signal detection measures of sensitivity and bias (*d'* and *criterion*) using the loglinear method to overcome extreme hit and false alarm rates of 1 or 0 (Hautus, 1995; see also, Stanislaw & Todorov, 1997), which were subsequently compared across the three tasks via two separate one-way ANOVAs. Analysis of *d'* showed an effect of task, $F(2,178) = 138.38$, $p < .001$, $\eta_p^2 = .61$. Tukey HSD revealed that this was due to lower sensitivity on the KFMT ($M = 0.71$, $SD = 0.58$) than the GFMT ($M = 1.98$, $SD = 0.99$) and the Avatar test ($M = 1.76$, $SD = 0.74$), both $ps < .001$, and sensitivity was higher for the GFMT than the Avatars, $p < .05$. Pairwise correlations on *d'* were found between the GFMT and KFMT, $r(88) = .62$, $p < .001$, the GFMT and Avatar test, $r(88) = .51$, $p < .001$, and the KFMT and the Avatar test, $r(88) = .54$, $p < .001$.

The corresponding analysis of *criterion* also revealed an effect of test, $F(2,178) = 6.69$, $p < .01$, $\eta_p^2 = .07$. Tukey's HSD shows that this was due to a more conservative *criterion* score on the KFMT ($M = 0.09$, $SD = 0.48$) in comparison with the GFMT ($M = -0.09$, $SD = 0.47$), $p < .01$, and Avatar test ($M = -0.11$, $SD = 0.69$), $p < .01$, reflecting a tendency to make fewer match responses on the KFMT. However, the difference in *criterion* between the GFMT and Avatar test was not significant, $p = .94$. Finally, *criterion* was compared to zero for all three tasks via a series of one-sample *t*-tests. These revealed that

remained significant following the application of this procedure, with the highest *p*-value ($p < .001$) surviving a critical threshold of .025. This analysis was not preregistered but included on the request of a reviewer.

criterion did not differ significantly from zero for any of the three tests, all *t*s ≤ 1.58, all *p*s ≥ .064.

Discussion

This experiment compared performance for two established laboratory tests of face matching with the identification of avatars. Converging with previous studies, accuracy was higher for the GFMT than the KFMT (Fysh & Bindemann, 2018). This reflects differences in the constructions of these tests. The identity match trials on the GFMT were recorded on the same day and under similar conditions to examine identification accuracy under optimised conditions. In contrast, the KFMT is constructed from a more varied face set in which the identity match trials were based on images that were acquired several months apart. Here, the construction of the avatar test was more comparable to the GFMT, by pairing face photographs with avatar face scans that were acquired on the same day, and performance for these two tests was similar. This suggests that the construction characteristics of laboratory face tests such as the GFMT are also preserved in the Avatar test, and transcend identification across these different mediums.

Of primary interest was whether associations in individual performance would also be found between laboratory face tests and the avatars. Such correlations emerged between all three tests on measures of overall accuracy and for match and mismatch trials. The strength of these correlations between the avatar and face tests was moderate, and is similar to those that were obtained when the GFMT and KFMT are compared directly. Thus, the data provide consistent evidence that identity matching of avatars is comparable to that of pairs of face photographs.

Experiment 2

Considering the correspondence between the identification of avatars and the face-matching tests in Experiment 1, it is critical to know whether these widely-used laboratory tests of face matching also relate to person identification when the avatars are presented in a more complex setting in VR. Experiment 2 investigates this question by presenting the avatars as travellers passing through passport control in a VR airport. We examined the accuracy with which these avatars were compared with a photo-ID document, and then compared this with performance on the GFMT and KFMT.

Method

This experiment was preregistered on OSF (https://osf.io/jcu74/?view_only=).

Participants

Sixty participants (33 females, 27 males) with a mean age of 34.6 years (*SD* = 10.6, range: 18-65) were recruited to take part in this study in exchange for a small fee using Prolific Academic. All participants resided in the United Kingdom at the time of testing, and none had participated in Experiment 1. An *a priori* power analysis suggested that a sample of 54.7 participants was sufficient to detect a correlation effect of $r = .46$ (mean of GFMT/KFMT vs. Avatar correlations in Experiment 1) with a statistical power level of .95 and an alpha threshold of $p = .05$. Because the closest integer divisible by six (the possible task order) is 60, this became our target sample size. This experiment was conducted live in a Zoom call with participants interacting with the experimenter, and so there were no attention check trials, and no participants were excluded from the final analysis.

Tasks

The GFMT and KFMT were constructed as in Experiment 1. In contrast, the avatars and their corresponding face photos were now presented in a passport control context in a VR airport. This passport control environment was constructed by positioning 3D objects within a pre-built 3D airport hall model (https://www.turbosquid.com/3d-models/airport-departures-lounge-3d-model/626226). This model was built in 3DS Max and used VRay for rendering. The completed passport control environment consisted of a booth area in which the participants were standing, equipped with a desk, chair and computer. This booth was situated inside the airport hall with other visual cues, such as departure boards, clearly visible to participants. This airport environment is illustrated in Figure 4 (for further details, see Tummon et al., 2019, 2020).

FIGURE 4. The upper panel provides an illustration of the airport environment in Experiment 2, from the perspective of a participant performing the identification task. The lower panel depicts a range of exemplar avatars that made up this task.

Procedure

This study was conducted during the Covid-19 global pandemic, preventing in-person testing. To overcome this issue, the three tasks that feature in this study were run on a remote computer and screen-shared with participants via telecommunications software (Zoom). Participants completed all three tasks (GFMT, KFMT, Airport) by providing verbal responses (i.e., 'same' / 'different'), which were then registered by the experimenter. For all tasks, accuracy of response was emphasised and, because data collection was conducted via telecommunications software, response times were not analysed. The order of the three tasks (GFMT, KFMT, Airport) was counterbalanced across participants.

The GFMT and KFMT were presented with *PsychoPy 3* software (Peirce, 2007). In these tests, each trial began with a 1-second fixation cross, followed by a face pair, which remained on display until a response was registered. The presentation of face pairs was randomised in both tests and participants were asked to categorize these stimuli as depicting the same person or different people.

The avatar identification task was presented using *Vizard 6* software. Participants saw a group of travellers arriving in the airport arrivals hall and forming a queue at the passport control desk. The avatars would wait in an idling mode, shifting slightly in stance to indicate a waiting body language. On each trial, an avatar would approach the participant situated in the passport control booth. A photo-ID card with a digital photograph would then appear next to the avatar, which would display either a photograph of the same person or of a different identity (see Figure 4). Observers classified each pairing verbally as depicting the same person or different people. The avatar then walked past the control booth on the left side if classified as an identity match or to the right if classified as a mismatch, and disappeared out of view thereby triggering the next avatar to approach passport control. The order of avatar presentation was randomised.

Results

Accuracy

The mean percentage of correct responses was calculated for match and mismatch trials for each test, and are illustrated in Figure 5. A 3 (test: GFMT, KFMT, Airport) x 2 (trial type: matches, mismatches) within-subjects ANOVA of these data revealed main effects of test, $F(2,118) = 105.68$, $p < .001$, $\eta_p^2 = .64$, and trial type, $F(1,59) = 9.91$, $p < .01$, $\eta_p^2 = .14$, and an interaction between these factors, $F(2,118) = 32.31$, $p < .001$, $\eta_p^2 = .35$. Analysis of simple effect tests showed an effect of task for match trials, $F(2,58) = 97.45$, $p < .001$, $\eta_p^2 = .77$. Tukey HSD test revealed that this was characterized by higher match accuracy on the Airport task than the GFMT and KFMT, both *ps* < .001, and on the GFMT than the KFMT, $p < .001$.

A simple main effect of task was also observed for mismatch trials, $F(2,58) = 37.80$, $p < .001$, $\eta_p^2 = .57$, whereby accuracy was again higher on the GFMT than the KFMT, $p < .001$. In contrast to match trials, however, accuracy on the GFMT also exceeded that for the Airport, $p < .001$, and did not differ significantly between the Airport and the KFMT, $p = .80$. Finally, analysis of simple main effects of trial type showed that match accuracy was greater than mismatch accuracy in the Airport, $F(1,59) = 55.20$, $p < .001$, $\eta_p^2 = .48$, however the difference in match and mismatch accuracy for both the GFMT and the KFMT was not significant, $F(1,59) = 0.23$, $p = .64$, $\eta_p^2 = .004$, and $F(1,59) = 0.48$, $p = .49$, $\eta_p^2 = .01$, respectively. Overall, this analysis shows that identification at the airport exceeded accuracy of the GFMT and KFMT on match trials. In contrast, mismatch accuracy for the avatars was similar to the KFMT, and lower than on the GFMT.

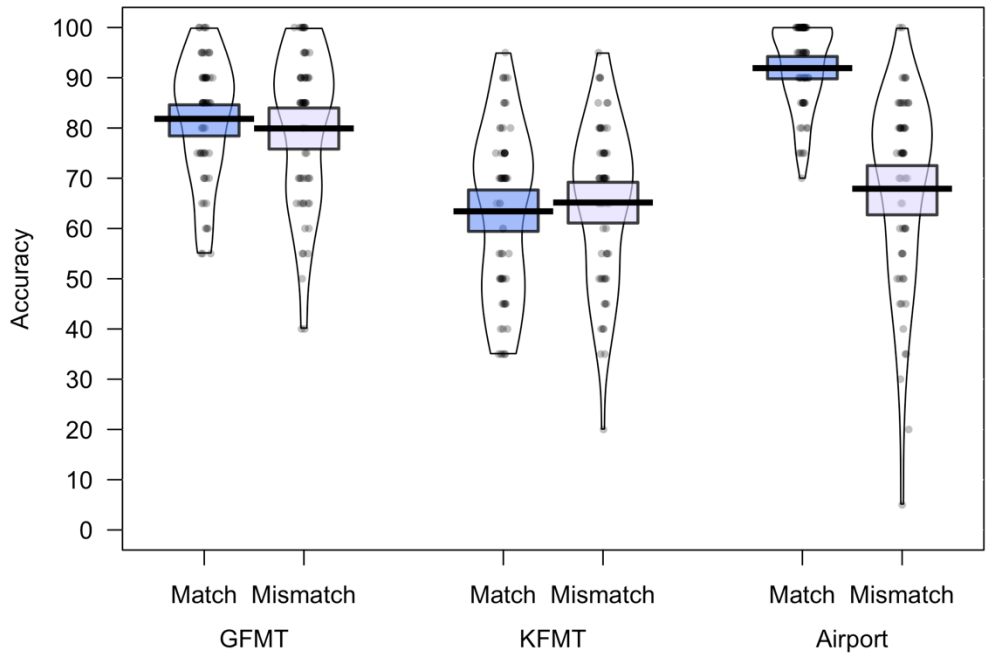

FIGURE 5. Accuracy (%) for match and mismatch trials across the three tasks employed in Experiment 2. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

To examine individual identification performance across the GFMT, KFMT and the Airport, pairwise Pearson's correlations were performed on overall, match and mismatch accuracy for all tests. These data are illustrated in Figure 6 and show positive correlations between the GFMT and KFMT in overall, match and mismatch accuracy, all $rs \geq .31$, $ps < .05$. Correlations were also observed between the GFMT and the Airport in overall, $r = .33$, $p < .05$, and mismatch accuracy, $r = .33$, $p < .05$, but not for identity matches, $r = .22$, $p = .09$. In turn, the KFMT correlated with the Airport in matches, $r = .37$, $p < .07$, and mismatches, $r = .44$, $p < .001$, but not in overall accuracy, $r = .12$, $p = .37$[5]. In summary, these data therefore show that the GFMT and KFMT correlated across all measures, whereas the correlations of these laboratory tests with the Airport were somewhat less consistent.

[5] In addition, all significant *p*-values survived the Benjamini-Hochberg correction, with the highest significant relationship ($p = .018$) surviving a critical threshold of .019.

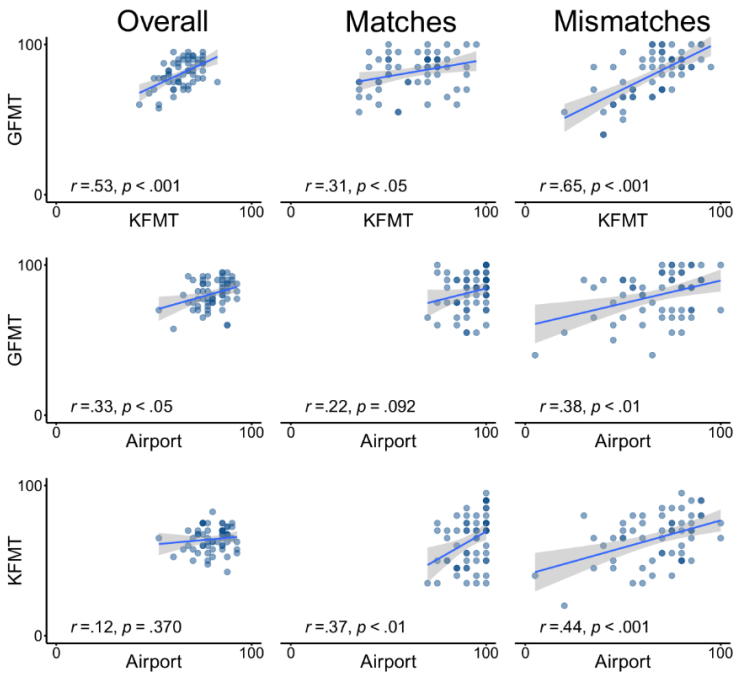

FIGURE 6. Accuracy (%) correlations between the GFMT, KFMT, and the Airport matching task for overall, match and mismatch accuracy in Experiment 2.

d' and criterion

Accuracy was also converted into *d'* and *criterion*. Analysis of *d'* showed an effect of task, $F(2,118) = 128.51$, $p < .001$, $\eta_p^2 = .69$, due to higher sensitivity for the Airport task ($M = 1.92$, $SD = 0.49$) than the KFMT ($M = 0.76$, $SD = 0.44$), $p < .001$, and on the GFMT ($M = 1.87$, $SD = 0.69$) compared to the KFMT, $p < .001$. In contrast, *d'* did not reliably differ between the GFMT and Airport task, $p = .82$. Positive correlations of *d'* were found between the GFMT and the KFMT, $r(58) = .47$, $p < .001$, and between the GFMT and the Airport, $r(58) = .37$, $p < .01$, but not between the KFMT and Airport task, $r(58) = .22$, $p = .10$.

ANOVA also revealed an effect of *criterion* across tasks, $F(2,118) = 39.25$, $p < .001$, $\eta_p^2 = .40$, due to a more liberal *criterion* on the Airport task ($M = -0.46$, $SD = 0.48$) than the KFMT ($M = 0.02$, $SD = 0.39$) and GFMT ($M = -0.01$, $SD = 0.42$), both $ps < .001$. This reflects a bias to make more match responses in the Airport than on the two face-matching

tests, for which *criterion* did not differ significantly, $p = .84$. Finally, one-sample *t*-tests revealed that *criterion* in the Airport task was below zero, $t(59) = 7.44$, $p < .001$, confirming a bias to make more *match* decisions, but did not differ from to zero on the GFMT, $t(59) = 0.26$, $p = .79$, and the KFMT, $t(59) = 0.40$, $p = .69$.

Discussion

The results for the laboratory tests of face matching converge closely with Experiment 1. Overall accuracy was higher on the GFMT than the KFMT, but correlations were observed consistently across all accuracy measures between these two tests. In contrast to Experiment 1, classification of avatar identity matches was enhanced, so that this now exceeded match accuracy on the GFMT, when these were presented in the airport. In turn, mismatch accuracy for avatars did not differ significantly from the KFMT, and was lower than on the GFMT. Some differences also emerged in correlations between experiments. In Experiment 1, for example, correlations were observed in every single measure between tests (GFMT, KFMT, Avatars). In Experiment 2, the correlations did not reach significance for the comparison of the GFMT and the Airport in identity matches, and the KFMT and Airport in overall accuracy. Moreover, correlations were generally weaker between the laboratory tests and the Airport in Experiment 2 (mean $r = .31$) than with the Avatars in Experiment 1 (mean $r = .46$).

Two insights emerge from these data. First, Experiment 2 demonstrates correspondence between laboratory paradigms of face matching and avatar identification, and extends these to a scenario where avatar identifications are made in the context of passport control at a VR airport. This indicates that VR can be used to conduct behaviourally-relevant research on face identification in more complex settings than traditional laboratory experiments allow, whilst maintaining correspondence between tasks. However, the data also

indicate that the airport context influenced behaviour in the face-matching task, as is evident from a bias to make more match responses in VR. This match response bias was characterised by near-ceiling accuracy on match trials in the airport task, and it is possible that the lack of variability caused by this was the main reason that an association was not observed for match accuracy between the GFMT and the Airport. The presence of a match response bias here also converges with other experiments, which suggest that contextual cues, such as photo-identity documents, increase false acceptance of identity mismatches (Feng & Burton, 2019, in press; McCaffery & Burton, 2016). This is an important finding, as it indicates that the environment or social context within which these identifications occur affects the task outcome.

Experiment 3

Experiment 2 demonstrates that VR can be used to conduct behaviourally-relevant research on face identification in more complex environments, whilst maintaining correspondence with laboratory tasks. In Experiment 3, we seek to strengthen these findings by assessing the consistency of face identification in VR. Compared to standard laboratory tests of face matching, in which observers compare pairs of isolated face images on a computer screen, person identification within the context of the VR airport setting introduces further variables into this task, such as the visibility of a passenger queue, avatar movement, visual objects, and airport scenery. This can create additional variation in participants' task experience that does not exist in laboratory tests and could affect identification in VR. It is therefore important to determine the consistency of identification in VR, considering its novelty to this research field. Experiment 3 therefore assesses test-retest reliability of identification in the airport, by recording performance across a delay of one week. We also expanded the duration of the airport test to encompass more trials and a more varied set of

faces. Once again, we compared identification at the Airport with performance on the GFMT and KFMT.

Method

This experiment was preregistered on OSF (https://osf.io/jcu74/?view_only=).

Participants

Sixty participants (31 females, 29 males) with a mean age of 33.6 years (*SD* = 9.3, range: 18-62) were recruited via Prolific Academic to participate in this study in exchange for a small fee. All participants resided in the United Kingdom at the time of recruitment. None of the participants for this experiment took part in Experiments 1 or 2. As for Experiment 2, an *a-priori* power analysis suggested that a sample of 54.7 participants was sufficient to detect a correlation effect of $r = .46$ with a statistical power level of 0.95 and an alpha threshold of $p = .05$. Because the closest integer divisible by six (the possible task order) is 60, this became our target sample size. In addition, as with Experiment 2, there were no attention check trials because this experiment was administered to participants via Zoom, and no participants were excluded from the final analysis.

Tasks and Procedure

The method and procedure of the GFMT and KFMT were identical to Experiment 2, but the Airport task was extended to encompass an additional 40 trials (20 matches and 20 mismatches) that were randomly intermixed with the original 40 trials featuring in Experiments 1 and 2, resulting in 80 trials total (40 matches, 40 mismatches). Unlike the GFMT, KFMT and the avatars that featured in the previous two experiments, these 40 new

trials portrayed a mixture of different ethnicities, to capture the many different nationalities of people that navigate through real-life airports (for examples, see Figure 7).

FIGURE 7. An illustration of the range of avatars encountered in the airport of Experiment 3.

Performance on all three tasks was measured repeatedly, by testing every participant twice, with an average interval of 7.1 days (*SD* = 0.5) between the first and second test session. All three tasks (including the face stimuli) were identical between testing sessions 1 and 2. The order of the GFMT, KFMT and Airport task was counterbalanced across participants over the course of the experiment, but was held constant for each participant for testing at time 1 and 2.

Results

Accuracy

The mean percentage of correct responses was calculated for match and mismatch trials for all tests and both test sessions, and are illustrated in Figure 8. A 3 (test: GFMT, KFMT, Airport) x 2 (trial type: matches, mismatches) x 2 (time: time 1, time 2) within-subjects ANOVA revealed main effects of task, $F(2,118) = 203.91$, $p < .001$, $\eta_p^2 = .78$, and of trial type, $F(1,59) = 8.14$, $p < .01$, $\eta_p^2 = .12$, and an interaction between these factors $F(2,118) = 28.28$, $p < .001$, $\eta_p^2 = .32$. Analysis of simple main effects revealed an effect of task for matches, $F(2,58) = 150.22$, $p < .001$, $\eta_p^2 = .84$. Tukey HSD test showed that match accuracy

was higher in the Airport than in the KFMT, $p < .001$, and the GFMT, $p < .05$. Likewise, match accuracy in the GFMT was also greater than for the KFMT, $p < .001$. A simple main effect of task was also found for mismatches, $F(2,58) = 21.67$, $p < .001$, $\eta_p^2 = .43$. Tukey HSD test showed that accuracy was higher on the GFMT than in the Airport, $p < .01$, and the KFMT, $p < .001$. Mismatch accuracy in the Airport was also greater than in the KFMT, $p < .01$. Finally, simple main effects of trial type were found for the GFMT, $F(1,59) = 6.36$, $p < .05$, $\eta_p^2 = .10$, and Airport, $F(1,59) = 35.87$, $p < .001$, $\eta_p^2 = .38$, due to higher accuracy on match compared to mismatch trials. In contrast, match accuracy did not differ significantly from mismatch accuracy in the KFMT, $F(1,59) = 0.62$, $p = .43$, $\eta_p^2 = .01$.

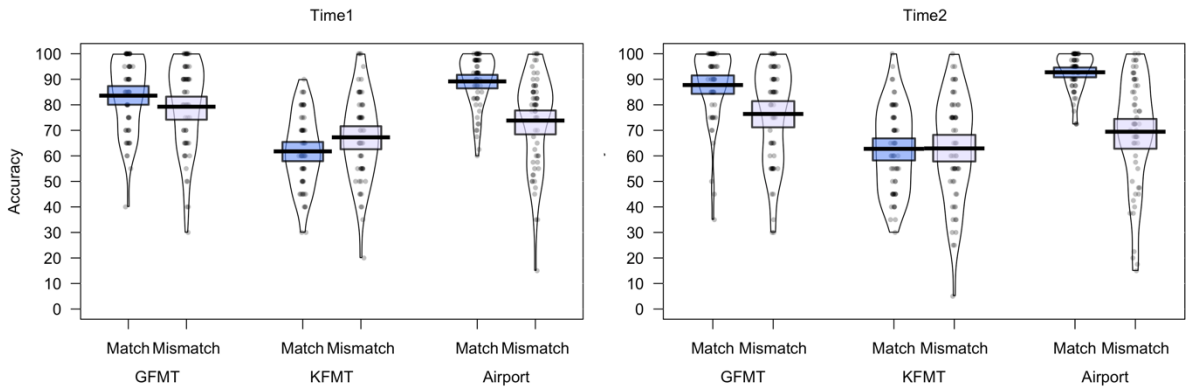

FIGURE 8. Accuracy on the GFMT, KFMT and Airport matching task in Experiment 3, for time 1 (left) and time 2 (right). The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

In addition, an interaction of trial type and time was found, $F(1,59) = 17.59$, $p < .001$, $\eta_p^2 = .23$. Analysis of simple main effects reveal an effect of trial type for time 2, $F(1,59) = 12.65$, $p < .01$, $\eta_p^2 = .18$, reflecting higher accuracy on match than mismatch trials, but not for time 1, $F(1,59) = 3.18$, $p = .08$, $\eta_p^2 = .05$. A simple main effect of time was also observed for matches, $F(1,59) = 10.28$, $p < .01$, $\eta_p^2 = .15$, due to higher accuracy at time 2 than at time 1, and for mismatches, $F(1,59) = 11.35$, $p < .01$, $\eta_p^2 = .16$, due to the reverse pattern. Finally, an

interaction between time and task, $F(2,118) = 1.73$, $p = .18$, $\eta_p^2 = .03$, and a three-way interaction were not found, $F(2,118) = 0.41$, $p = .67$, $\eta_p^2 = .01$.

Overall, these data show that match accuracy on the airport exceeded that of the GFMT and KFMT. In contrast mismatch accuracy at the airport was similar to the KFMT and lower than on the GFMT. In addition, match accuracy also increased generally, and mismatch accuracy decreased, across test sessions.

To examine individual identification performance, pairwise Pearson's correlations were performed on accuracy for all tests. These correlations are illustrated in Figure 9 and revealed consistent positive associations between the GFMT, KFMT and Airport in overall, match and mismatch accuracy. This pattern was observed at time 1, all $rs \geq .31$, $ps < .05$, and also at time 2, all $rs \geq .40$, $ps < .01$.

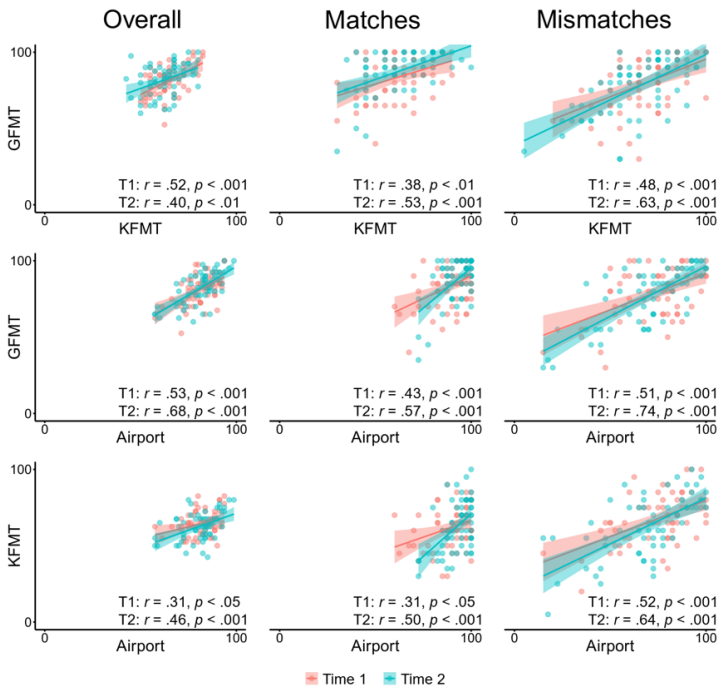

FIGURE 9. Accuracy (%) correlations between the GFMT, KFMT and the Airport matching task in Experiment 3 at time 1 (red) and time 2 (green).

In addition, we also examined test-retest reliability, by correlating performance across time and time 2. These data are illustrated in Figure 10 and show that correlations were observed for all tests, and on all matching measures, across time 1 and time 2, all $rs \geq .50$, $ps < .001$[6].

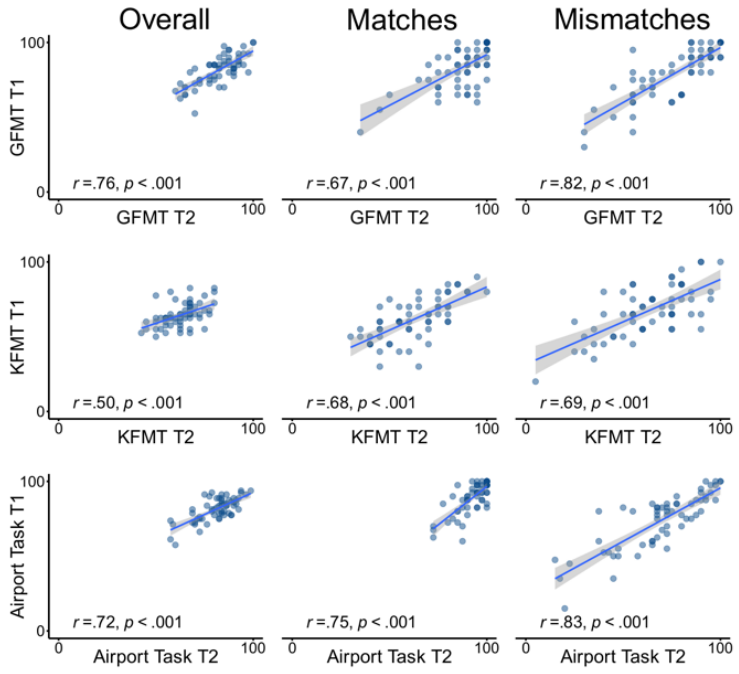

FIGURE 10. Test-retest correlations, comparing time 1 (T1) and time 2 (T2), for the GFMT, KFMT and Airport in overall, match and mismatch accuracy.

d' and criterion

Accuracy was also converted into *d'* and *criterion*, which were analysed via separate 2 (time: time 1, time 2) x 3 (task: Airport, KFMT, GFMT) within-subject ANOVAs. For *d'*, this analysis showed a main effect of task, $F(2,118) = 218.15$, $p < .001$, $\eta_p^2 = .79$, due to higher sensitivity on the GFMT ($M = 2.04$, $SD = 0.79$) than the KFMT ($M = 0.75$, $SD = 0.52$), $p < .001$, and on the Airport task ($M = 2.14$, $SD = 0.60$) compared to the KFMT, $p <$

[6] As in Experiment 1 and 2, the Benjamini-Hochberg procedure was applied to guard against false positive correlations. Following this procedure, all correlations remained significant with the highest *p*-value (.017) falling below the critical threshold (.025).

.001. In addition, sensitivity did not differ significantly between the GFMT and the Airport task, $p = .32$. There was no effect of time, with numerically similar rates of *d'* between time 1 ($M = 1.63$, $SD = 0.49$) and time 2 ($M = 1.66$, $SD = 0.53$), $F(1,59) = 0.39$, $p = .54$, $\eta_p^2 = .01$, and no interaction between factors, $F(2,118) = 1.95$, $p = .15$, $\eta_p^2 = .03$.

The equivalent analysis of *criterion* revealed an effect of time, $F(1,59) = 20.98$, $p < .001$, $\eta_p^2 = .26$, due to a greater tendency to submit identity match responses at time 2 ($M = -0.23$, $SD = 0.44$) than in time 1 ($M = -0.10$, $SD = 0.37$). An effect of task was also present, $F(2,118) = 29.59$ $p < .001$, $\eta_p^2 = .33$, reflecting a more liberal response criterion whereby faces were more likely to be classified as identity matches in the Airport ($M = -0.39$, $SD = 0.55$) than in the KFMT ($M = 0.05$, $SD = 0.44$) and GFMT ($M = -0.15$, $SD = 0.50$), both ps $< .001$. Observers were also more inclined towards match responses in the GFMT than in the KFMT, $p < .01$. The interaction between task and time was not significant, $F(2,118) = 1.22$, $p = .30$, $\eta_p^2 = .02$.

Separate one-sample *t*-tests showed that *criterion* was reliably below zero for the Airport task at time 1, $t(59) = 4.38$, $p < .001$, and time 2, $t(59) = 6.73$, $p < .001$. For the GFMT, *criterion* did not differ significantly from zero at time 1, $t(59) = 1.26$, $p = .21$, but was significantly below zero at time 2, $t(59) = -3.33$, $p < .01$. Finally, for the KFMT *criterion* did not differ significantly from zero at both time 1, $t(59) = 1.82$, $p = .07$, and time 2, $t(59) = 0.08$, $p = .94$. Thus, identifications at the Airport were marked by a bias to make match responses, whereas a similar bias was only present in one of the laboratory tests (GFMT), and only at time 2.

Discussion

This experiment extends the key findings of Experiments 2 to a longer version of the Airport with more varied avatars. Matching performance correlated for the face tests and the

Airport, indicating correspondence between laboratory tasks and person identification in VR. As in Experiment 2, however, performance at the airport was marked by a response bias, whereby observers were more likely to classify face pairings as matches in this environment. Experiment 3 extends these findings in an important way, by demonstrating the same pattern of responses across two separate testing sessions, which were recorded one week apart. Correlation of performance in the airport at time 1 and time 2 was high across all measures (all *rs* > .7). This indicates good test-retest reliability for avatar identification in the airport, despite the additional variance that the VR environment can introduce into participants' experience. This shows that VR can provide a consistent and dependable method for studying person identification in more complex environments.

Finally, Experiment 3 also revealed a response bias, whereby observers were more likely to make match decisions at time 2 than time 1, irrespective of task. A similar bias has been observed in studies that have examined face matching over prolonged testing sessions (e.g., Alenezi & Bindemann, 2013; Fysh & Bindemann, 2017; Papesh et al., 2018), and which persists even when 5-minute rest breaks are introduced throughout the task (Alenezi et al., 2015). Although we cannot resolve the basis of this bias here, Experiment 3 extends such findings by indicating that this effect is also not mitigated by a task interval of several days.

General Discussion

Psychological experiments typically trade off tight control of variables in the laboratory against the ecological validity of field research. In recent years, the development of immersive VR has emerged as a possible link between these approaches, by providing complex environments for studying human cognition that also preserve control over key components of a study. However, investigations into the correspondence of VR to the traditional laboratory approach are still limited. In this study, we investigated this

correspondence by focusing on the identity matching of faces. We compared two established face-matching tests from the laboratory (GFMT, KFMT), in which pairs of faces are presented in simple displays, against the identification of person-avatars at passport control in a VR airport.

Performance for the laboratory tests converged with previous work, demonstrating higher accuracy on the GFMT than KFMT, and positive correlations between these tests (Fysh & Bindemann, 2018). A similar pattern emerged when these tests were compared with avatar faces that were presented in laboratory-style displays in Experiment 1, whereby the matching of these avatars to face photographs was comparable to the GFMT, and more accurate than for the KFMT. We attribute this finding to the construction of the avatar test, which was more similar to the GFMT, by pairing face photographs with face scans that were acquired on the same day. This suggests that the construction characteristics of laboratory face tests such as the GFMT are preserved in the avatar test, and transcend identification across these different mediums. In addition, similar accuracy correlations were observed between the identification of avatars and the face tests.

A similar pattern emerged when the avatars were encountered as animated travellers at passport control in a VR airport (Experiment 2 and 3). Under these conditions, correlations between laboratory tests and avatar identification persisted, indicating that individuals who performed well with the controlled face-matching tests were also more likely to accurately match people in the airport. Indeed, while a small number of correlations failed to reach significance (see Experiment 2), and broad individual differences in performance were observed in all tasks, the general pattern of correlations was remarkably similar across experiments (c.f., Figures 3, 6 and 9). These individual differences are consistent with other face-matching studies, which have shown similar variation in novices (e.g., Burton et al., 2010; Fysh & Bindemann, 2018) and practitioner groups (e.g., Phillips et al., 2018; White et

al., 2014; see Lander et al., 2018, for a review), while the magnitude of correlations between tests also converges with previous work (e.g., Burton et al., 2010; Fysh & Bindemann, 2018; Tummon et al., 2019, 2020). In addition, face matching in the GFMT, KFMT and airport also showed similar test-retest reliability. Taken together, these findings suggest that person identification in the VR airport environment corresponds with face identification in the laboratory tests.

However, the airport context also appeared to change the demands of the identification task. Performance for match and mismatch trials was comparable in the GFMT and KFMT throughout this study, and this was also found in the avatar task of Experiment 1. When the same avatars were presented in the airport in Experiments 2 and 3, a response bias emerged whereby observers were more likely to falsely accept identity mismatches. This indicates that there is correspondence between face matching in simplified laboratory tests and more complex contexts, but also some differences that qualify performance in this task. These findings converge with other research which indicates that contextual manipulations that mimic specific aspects of applied settings produce similar response biases in person identification. The practice of embedding photographs in photo-identity documents, for example, also biases face identification by impairing classification of mismatches (Feng & Burton, 2019; McCaffery & Burton, 2016). Furthermore, a study of passport officers making photo-ID checks on live participants also showed a bias towards accepting more 'fraudulent' ID, i.e., mismatch trials (White et al., 2014).

The difference between face identification in the laboratory tasks and at the virtual airport speaks to the importance of studying face identification in contexts which mimic the nature of the setting in which this task is actually performed. Our experiments stop short of systematically investigating the various contextual factors that might affect identification in the airport task, such as the visibility of a passenger queue, avatar movement, or different

components of the airport scenery. As a consequence, it is difficult to disentangle whether the response bias in Experiments 2 and 3 should be attributed to the identification of moving avatars (but see Tummon et al., 2020), the task context, or both. This issue should be addressed in future experiments with VR. In such work, it may also be possible to depict avatars alongside face photographs of the same person that were acquired several months earlier or later, which might lead the matching of avatars to emulate the characteristics of the KFMT more closely than it did here. This would increase the correspondence of our VR airport to real-world airport settings still further.

It is easy to understand why complex real-world factors are often ignored in face research, because it is difficult to implement environmental and social contexts with traditional experimental paradigms. In turn, these research questions are not easily addressed in the field. The study of person identification at the passport control of a *real* airport is difficult to achieve, for example, because this requires access to security-sensitive operational environments. Though it is sometimes possible to create field experiments, their operational complexity tends to render them 'one-shot' studies that are practically impossible to replicate – hence limiting the generalisability of results. The current approach provides a sophisticated method for simulating these settings instead and, in contrast to field settings, also retains experimental control and measurement. We also show that this approach provides good test-retest reliability, to demonstrate that VR can deliver good reproducibility, too.

In doing so, the current experiments provide an important demonstration of the behavioural relevance of VR to the psychological study of face perception. The simplified laboratory paradigms that are prevalent in the study of human cognition provide a limited context for understanding behaviour in the social and environmental contexts within which it occurs. It is becoming increasingly clear that these contextual factors cannot be regarded as simple 'add-ons' to understanding the process of face perception, but rather, represent an

important component in the processes of perception (Cole et al., 2016; Hayward et al., 2017; Ramon et al., 2019; Skarratt et al., 2012). The current study demonstrates that, through VR, it is possible to link the study of controlled laboratory processes and the complex contexts in which these processes typically occur.

Acknowledgements

This work was funded by a research grant from the Economic and Social Research Council (ESRC Grant no ES/S010181/1) to Markus Bindemann, A. Mike Burton and Cade McCall.

Author contributions

MB, MCF, IVT, CM, AMB conceived and designed the studies; MCF, IVT collected the data; MB, MCF, IVT, CM, AMB analysed and interpreted the data; MB, MCF, IVT, CM, AMB wrote the manuscript; JA constructed the VR airport.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-407. doi:10.3758/s13428-019-01237-x

Bailenson, J., Beall, A., & Blascovich, J. (2003). Using virtual heads for person identification: An empirical study comparing photographs to photogrammetrically-generated models. *Journal of Forensic Identification, 53*, 722-728.

Bailenson, J., Davies, A., Blascovich, J. J., Beall, A. C., McCall, C., & Guadagno, R. E. (2008). The effects of witness viewpoint distance, angle, and choice on eyewitness accuracy in police lineups conducted in immersive virtual environments. *Presence: Teleoperators and Virtual Environments, 17*, 242-255. doi:10.1162/pres.17.3.242

Bindemann, M. (2021). *Forensic Face Matching: Research and Practice*. Oxford University Press. ISBN 978–0–19–883774–9

Bindemann, M., & Megreya, A. M. (2017). *Face processing: Systems, Disorders and Cultural Differences*. New York: Nova Science Publishers, Inc. ISBN: 978-1-53612-398-2

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

Bruce, V., & Young, A. W. (1998). *In the eye of the beholder: The science of face perception*. Oxford University Press.

Bülthoff, I., Mohler, B. J., & Thornton, I. M. (2019). Face recognition of full-bodied avatars by active observers in a virtual environment. *Vision Research, 157*, 242–251. doi:10.1016/j.visres.2017.12.001

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

Cole, G. G., Skarratt, P. A., & Kuhn, G. (2016). Real person interaction in visual attention research. *European Psychologist, 21*, 141-149. doi:10.1027/1016-9040/a000243

Cousineau, D. (2005). Confidence intervals in within-subjects designs: A simpler solution to Loftus and Masson's method. *Tutorial in Quantitative Methods for Psychology, 1*, 42-45. doi:10.20982/tqmp.01.1.p042

Feng, X., & Burton, A. M. (2019). Identity documents bias face matching. *Perception*, *48*, 1163-1174. doi:10.1177/0301006619877821

Feng, X., & Burton, A. M. (in press). Understanding the document bias in face matching. *Quarterly Journal of Experimental Psychology*. doi:10.1177/17470218211017902.

Fysh, M. C. (2021). *Factors limiting face matching at passport control and in police investigations*. In Markus Bindemann (ed.), Forensic Face Matching: Research and Practice. (pp. 255-261) Oxford University Press, UK. doi:10.1093/oso/9780198837749.003.0011

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face matching accuracy. *Royal Society Open Science, 4*:170249, 1-13. doi:10.1098/rsos.170249

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology, 109,* 219-231. doi:10.1111/bjop.12260

Fysh, M. C., Trifonova, I. V., Allen, J., McCall, C., Burton, A. M., & Bindemann, M. (in press). Avatars with Faces of Real People: A Construction Method for Scientific Experiments in Virtual Reality. *Behavior Research Methods.*

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*, 46-51. doi:10.3758/BF03203619

Hayward, D. A., Voorhies, W., Morris, J. L., Capozzi, F., & Ristic, J. (2017). Staring reality in the face: A comparison of social attention across laboratory and real-world measures suggests little common ground. *Canadian Journal of Experimental Psychology*, *71*, 212-225. doi:10.1037/cep0000117

Hole, G. J., & Bourne, V. (2010). *Face processing: Psychological, neuropsychological, and applied perspectives*. Oxford University Press.

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*:26, 1-13. doi:10.1186/s41235-018-0115-6

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers*, *31*, 557–564. https://doi.org/10.3758/BF03200735

McCaffery, J. J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications, 3*:21, 1-15. doi:10.1186/s41235-018-0112-9

McCall, C., & Blascovich, J. J. (2009). How, when, and why to use digital experimental virtual environments to study social behavior. *Social and Personality Psychology Compass*, *3*, 744–758. doi:10.1111/j.1751-9004.2009.00195.x

McCall, C., Hildebrandt, L. K., Hartmann, R., Baczkowski, B. M., & Singer, T. (2016). Introducing the Wunderkammer as a tool for emotion research: Unconstrained gaze and movement patterns in three emotionally evocative virtual worlds. *Computers in Human Behavior, 59*, 93-107. doi:10.1016/j.chb.2016.01.028

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition

algorithms. *Proceedings of the National Academy of Sciences, 115*, 6171-6176. doi:10.1073/pnas.1721355115

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology, 110*, 461-479. doi:10.1111/bjop.12368

Rhodes, G., Calder, A. J., Johnson, M., & Haxby, J. V. (2011). *Oxford handbook of face perception*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199559053.001.0001

Skarratt, P., Cole, G. G., & Kuhn, G. (2012). Visual cognition during real social interaction. *Frontiers in Human Neuroscience, 6,* 42979. doi:10.3389/fnhum.2012.00196

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137-149. doi:10.3758/BF03207704

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., et al. (2019). Do professional facial image comparison training courses work? PLoS ONE, 14:2, e0211037, 1-17. doi:10.1371/journal.pone.0211037

Tummon, H. M., Allen, J., & Bindemann, M. (2019). Facial identification at a virtual reality airport. *i-Perception*, *10*, 2041669519863077. doi:10.1177/2041669519863077

Tummon, H. M., Allen, J., & Bindemann, M. (2020). Body language influences on facial identification at passport control: an exploration in virtual reality. *i-Perception*, *11*, 2041669520958033. doi:10.1177/2041669520958033

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE, 10*:10, e0139827, 1-14. doi:10.1371/journal.pone.0139827

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

Wilson, C. J., & Soranzo, A. (2015). The use of virtual reality in psychology: A case study in visual perception. *Computational and Mathematical Methods in Medicine*, 1–7. https://doi.org/10.1155/2015/151702

Wirth, B. E. & Carbon, C. C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied, 23*, 138-157. doi:10.1037/xap0000114