Article:

McCaffery, Jennifer Maureen orcid.org/0000-0003-4420-290X and Burton, Anthony Michael orcid.org/0000-0002-2035-2084 (2016) Passport checks:interactions between matching faces and biographical details. Applied Cognitive Psychology. pp. 925-933. ISSN: 0888-4080

https://doi.org/10.1002/acp.3281

Passport checks: interactions between matching faces and biographical details

Journal:	*Applied Cognitive Psychology*
Manuscript ID	ACP-16-0049.R1
Wiley - Manuscript type:	Research Article
Keywords:	face recognition, face matching, passports

Abstract

Matching unfamiliar faces is known to be a difficult task. However, most research has tested viewers' ability to match pairs of faces presented in isolation. In real settings, professionals are commonly required to examine photo-ID which contains other biographical information too. In three experiments we present faces embedded in passport frames, and ask viewers to make face matching decisions, and to check biographical information. We find that the inclusion of a passport frame reduces viewers' ability to detect a face mismatch. Furthermore, the nature of the face-match influences biographical-data checking – true matches lead to fewer detections of invalid data. In general, viewers were poor at spotting errors in biographical information. This pattern suggests that detection of fraudulent photo-ID is even harder than current experimental studies suggest. Possible mechanisms for these effects are discussed.

Introduction

It is now well-established that unfamiliar face matching is a difficult task, on which viewers are prone to errors (e.g., Bruce et al, 1999; Johnston & Edmonds, 2009; Megreya & Burton, 2006). This difficulty is typically demonstrated by lab-based experiments in which viewers are asked to make same/different decisions to two photographs, however it is also observed when matching a live person to a photo (e.g. Davis & Valentine, 2009; Kemp, Towell & Pike, 1997; Megreya & Burton, 2008). Furthermore, expert observers, including passport officers, are typically no more accurate than the general population (White, Kemp, Jenkins, Matheson & Burton, 2014).

While these findings appear to be important for informing practical uses of face recognition, the experiments typically do not capture an important aspect of day to day experience with checking photo ID. In reality, viewers rarely see faces in isolation (Heyer & Semmler, 2013). Instead, photos are usually embedded in documents such as passports, driving licences, or other photo-ID. These documents usually contain important information about the bearer (e.g. name, age, address), and people checking ID are often required to confirm some of these details - for example age-checks for alcohol sales. This raises two questions: first, does the presence of biographical data within a document context make any difference to people's face matching ability; second, does the presence of faces affect a viewer's ability to check the biographical data?

Although most laboratory-based face recognition work has used images of isolated faces, some previous experiments have presented these in the context of a real document. Kemp et al (1997) asked volunteer shoppers to show a mocked-up payment card to cashiers, while Bindemann & Sandford (2011) showed participants photos of university identity cards. However, neither of these experiments required participants to check the data shown on these cards, and neither provide a direct comparison between seeing faces within a document context and seeing them alone. In the following experiments, we therefore set out to test whether these factors influence one's ability to match faces. The intention is to provide an important missing link between laboratory-based face matching experiments, and the demands of day to day operation by professionals checking ID.

There are some good reasons to hypothesise that there may be an interaction between face matching and checking document data. First, any observer required to make *both* a judgement on a face, and a judgement about the likely veracity of document information is under greater cognitive load than someone simply required to make a judgement about faces. Of course, taskflow could be designed to attempt to optimise viewers' performance, but in a typical ID card (e.g. passport), information about the carrier's face and personal information is simultaneously present and so it is plausible that the simple presence of multiple information sources is distracting. Second, there is the possibility of interference between the two types of information. Kassin, Dror & Kukucka (2013) provide a thorough review of the various sources of 'forensic confirmation bias' in image comparison or memory, which can affect lay viewers and experts alike. Given the potential sources of bias, might a false piece of information on a document (for example one which indicates the wrong sex or age), lead to a bias to reject two face images as being the same person?

These issues have their source in theoretical understanding of human information processing, which consistently show that tasks with multiple demands are harder than single-demand processes (for a forensic example, see Menneer et al, 2012). Furthermore, it is well-established that the presence of faces can interfere with other perceptual tasks, even when the faces are task irrelevant (e.g. Jenkins, Lavie & Driver, 2003; Lavie, Ro & Russell, 2003). However, in addition to its theoretical interest, this is also a practical problem across a variety of settings. For example, border control is not the only situation in which face matching within a document takes place. In many countries passport renewal authorities attempt to guard against fraud in multiple ways, including comparison of application photos to previous applications or other official documentation (e.g. driving licences). The most recent available data for our own jurisdiction, the UK, show that in the financial year 2013-14, 5.7 million passport applications were processed, and of these 0.15% (over eight thousand) were detected to be fraudulent (HM Passport Office, 2014). So, it is important to establish whether the typical lab-based face matching study, in which faces are shown in isolation, generalises well to a situation in which faces are seen within documents.

In the following experiments we present viewers with pairs of faces, sometimes embedding one of these images in a passport frame. We ask our participants both to judge whether the two photos show the same or different people, and to decide whether the personal data is accurate. To anticipate the results, we consistently find that the presence of a passport frame biases viewers to make 'same face' judgements – a bias which leads to acceptance of fraudulent documents. We also find generally poor levels of performance in spotting invalid biographical details, and this is further weakened by the presence of a face match.

Experiment 1

Introduction

In this first experiment we investigate matching a face to an ID document (a passport) containing biographical data. We are concerned with two issues. First, does embedding a face image in an ID document make it any easier or harder to match? Second, what are the effects of adding a biographical data check? To answer these questions we used stimuli from a standardized test of face matching, the Glasgow Face Matching Test (GFMT, Short Version; Burton. White & McNeill, 2010). This comprises 40 face image pairs, for each of which viewers decide 'same person' or 'different person'. Using these stimuli we constructed items in which one of the face pair was embedded into a stylized UK Passport (see Figure 1). We then asked participants to make a face match (same/different person) and also to check the data on the passport for accuracy. We are particularly interested here in the detection of potentially fraudulent identity documents. We therefore aim to establish conditions in which it is easy or difficult to spot somebody using the 'wrong' ID – i.e. the document of someone whose face is not their own.

Method

FIGURE 1 HERE PLEASE

Stimuli

The 40 face pairs from the GFMT were used as stimuli. In order to construct ID-document items, each of the test pairs was recreated in a version with the left-face embedded in a passport frame, and another with the right face embedded in a passport frame. Across all items, distance between the pair was kept constant. Biographical information (all fictitious) was designed to be valid or invalid with the associated face. Key personal data could be rendered invalid as follows: wrong-gender forename (see bottom item in Figure 1); nouns as forename (e.g. 'Fork'); unlikely ethnicity surname (e.g. the Sri Lankan name 'Selvaratnam' for a male Caucasian face); and wrong birth date (either an impossible date, e.g. 30 Feb; or a birth year more than 20 years discrepant from the age of the target faces). Errors such as these are commonly found in forged documentation, due to inexpert transcription, and passport granting authorities routinely check these details. Invalid passport frames contained only one of these errors.

Participants

Participants were 80 members (31 male) of an online experimental volunteer community (UK Qualtrics panel), who were over 18 and reported normal or corrected to normal vision (mean age = 39.86, range = 18-75). All were British Citizens who had lived in the UK for at least the last 10 years. Participants were reimbursed a small fee for their time.

Design and Procedure

This experiment was conducted on-line. Each participant saw 10 face-only pairs and 30 pairs embedded in passport frames (half left, half right, see Figure 1), with order of presentation randomized throughout (i.e. not blocked by condition). Half the items showed same-person pairs, and half different person pairs. For those pairs with passport frames, one third of the items showed invalid data (i.e. ten items per participant). Face pairs were counterbalanced across the experiment such that each pair occurred equally often in each condition. For face-only pairs, participants were asked, on-screen, '*Are the images of the same individual?*' and selected responses 'Same ' or 'Different' with a mouse. For pairs including a passport frame, participants were asked two questions: '*Is the data correct?*' and '*Are the images of the same individual*?', and made their responses with a mouse. Order of questions was counter-balanced, and all stimuli remained on the screen until responses had been made – i.e. until after both responses for the pairs including a passport frame.

Prior to the experiment participants were given two practice trials, a face-only pair, and a passport frame pair. They were asked to practise same/different face responses by selecting the appropriate button on screen. For the passport item, they were asked to judge whether the personal information was correct with respect to the person shown in the passport. Instructions for this decision were as follows: '*If the data matches the image on the passport, e.g. correct gender first name and year of birth click 'Yes'. If the data is factually incorrect or appears to contradict the image, click 'No'*.

Results

For passport frame conditions, the order in which participants were asked 'face' and 'data' questions was included in an initial analysis. However, this showed no main effect, and no interactions with other factors. We therefore collapse over the two different orders in the presentation of data below.

Face Matching Accuracy

Mean face matching accuracy is shown in Figure 2, with data broken down into match and mismatch trials. A 3 (presentation type: faces only, passport-valid and passport-invalid) x 2 (match type: same/different) within subjects ANOVA showed no main effect of presentation type ($F(2,158) = .84$, $p > .05$, $\eta_\rho^2 < .01$) or match type ($F(1,79) = 3.52$ $p > .05$, $\eta_\rho^2 = .04$) but there was a significant interaction ($F(2,158) = 13.49$, $p < .001$, $\eta_\rho^2 = .15$).

FIGURE 2 HERE PLEASE

Simple main effects analyses showed that presentation type affected both same and different match items ($F(2, 316) = 6.33$, $p < .005$, $\eta_\rho^2 = .04$); and ($F(2, 316) = 6.98$, $p < .001$, $\eta_\rho^2 = .04$), respectively. The presence of a passport frame leads to significantly poorer performance in the face matching task for ‘different’ items but improved performance for ‘same’ items. This pattern of data suggests that the presence of a passport frame introduces a bias to respond ‘same’ to face pairs – a bias which will lead to increased errors in accepting fraudulent passports.

Data Checking Accuracy

Mean data checking accuracy scores are shown in Figure 3, broken down into trials in which the face pairs showed the same or different identities. A 2 (data type: valid/invalid) x 2 (match type: same/different) within subjects ANOVA showed a significant main effect of data type ($F(1,79) = 104.58$, $p < .001$, $\eta_\rho^2 = .57$) , and a significant effect of match type ($F(1,79) = 4.23$, $p < .05$), $\eta_\rho^2 = .05$). This was qualified by an interaction ($F(1,79) = 6.13$, $p < .01$, $\eta_\rho^2 = .07$) which arises because performance on valid data is better when faces match than when they do not – a pattern which is absent for invalid data, (simple main effects: ($F(1, 158) = 10.34$, $p < .005$, $\eta_\rho^2 = .06$); and ($F(1, 158) = 0.55$, $p > .05$, $\eta_\rho^2 = .00$), respectively).

FIGURE 3 HERE PLEASE

In summary, participants were very poor at spotting invalid information – detecting it on less than 50% of occasions. There is also an influence of the (task-irrelevant) face match on performance with data checking. Participants are more likely to claim erroneously that there is a data error when the faces mismatch.

Discussion

This study shows a number of interesting effects. Across the two rather different tasks (face matching and data verification), our participants were poor at spotting 'fraudulent' passport use. For face matching, viewers were poorer at detecting a mismatch when one of the faces was embedded in a passport frame. By comparison to the procedure which is normally used in lab-based studies of face matching, this particular error throws up a potentially important area of concern. It is possible that the additional visual context makes the matching process more difficult, and particularly in ways which lead to a 'same face' response bias. Note that the problem arises equally for valid and invalid data alike, i.e. it does not seem that the data match itself leads to a bias in face matching performance. Instead, the mere presence of a passport frame appears to make the match harder.

Second, there is a very strong tendency to miss invalid ('fraudulent') data on the passport itself. Participants knew that they would be seeing some invalid data, and yet they failed to spot this on fewer than half the occasions it was present. We also observed a modifying effect of the faces on data-checking accuracy. When the faces did not match, viewers were more likely to claim a data error on a valid passport. This pattern of data seems to suggest that mismatching faces have an effect on data checking, even though the faces are task-irrelevant – detecting that the faces are different seems to generalise (falsely) to biographical data when it is in fact valid.

In general, these results show that matching a face to an ID-document is difficult. It is well-known that unfamiliar face matching is a hard task: embedding photos in an ID document makes it even harder. Furthermore this seems to be a systematic effect - any passport frame context (whether bearing correct or incorrect biographical data) tends to bias viewers to respond 'same' more often than is the case for simple face-only matching. Note that this is a risky bias – it leads to increased levels of acceptance of documents which should be rejected as fraudulent.

It is a little more difficult to be clear about the effects on data matching in this experiment. Although there are clearly poor levels of fraud detection within these documents, the erroneous data was inserted in an unsystematic way. There were a number of different types of information which could be wrong in the 'invalid' passports, and it is quite possible that some of these are easier to spot than others. Furthermore, we have no baseline data on the levels of accuracy which can be achieved in data matching when there is not a two-face matching task present at the same time. So, in the following experiment, we examine data checking accuracy independently of face matching – attempting to capture the task of checking a passport for internal consistency, rather than comparing it to a person. We also make a systematic examination of the different types of valid and invalid data available.

Experiment 2

Introduction

This experiment examines viewers' ability to check three types of information on a photo-ID document: gender, year of birth and place of birth. Research into the perception of age and gender has shown that individuals are highly skilled at both. Gender classification from face images is typically near ceiling (e.g. Bruce, Ellis, Gibling, & Young, 1987; O'Toole, Deffenbacher, Valentin, McKee, Huff, & Abdi, 1998). Whereas viewers can typically judge an adult's age from a photo to within five years (Moyse, 2014; Voelkle, Ebner, Lindenberger & Riediger, 2012). Anastasi & Rhodes (2006) showed that young adults could sort photographs into three age ranges (18–25 years, 35–45 years, 55–75 years) with a correct response rate of 83.1%. These findings suggest that any mismatch between the perceived age or gender of a face and the accompanying biographical data could be relatively easily identified.

Unlike gender and age, place of birth is not visually derivable from an unfamiliar face so error checking is entirely based on transcription accuracy. Errors in the spelling of place of birth can indicate a forged passport, particularly for unsophisticated copies by non-native English speakers (Fender, 2008; Leslie & Thimke, 1986; our interactions with the UK Passport Office reveal that such transcription errors are a trigger for detection of fraudulent documents). In the following experiment we manipulate errors in each of these three types of information, aiming to establish their detectability.

Method

Stimuli

96 faces (half male) were used from the GFMT (long form), i.e. the same database as for experiment 1. We constructed UK passport frames for each of these faces (see Figure 1). Biographical information was designed to be valid or invalid for the associated face. For invalid gender frames, a wrong-gender forename was inserted, and the sex data (M/F) was incorrect. For the invalid year of birth (YOB) frames, a date was inserted which would indicate an age 20 years older than the face on the passport. (NB the GFMT face database includes the age of each person when the photo was taken, and so this information is available). For invalid place of birth (POB) fames, a misspelled UK town was used, e.g. 'Luuton' rather than 'Luton'. In all cases there was only one data error in each frame.

Participants

Participants were 36 members (16 male) of an experimental volunteer community at the University of York consisting of students, staff and residents. All were over 18 and reported normal or corrected to normal vision (mean age = 24.83, range = 18-69) and did not have dyslexia. To ensure their data checking was not compromised by a lack of familiarity with place names, participants were also required to be British Citizens who spoke English as their native language. Participants were reimbursed a small fee for their time.

Design and Procedure

This experiment was conducted in the lab using E-Prime. Participants received instructions as follows: *'you will be shown a passport frame containing biographical data and a photo of an individual. The data in each passport frame should be checked for errors, e.g. wrong gender, poor spelling or an unlikely age match between the date of birth and estimated age from photo. Click 'correct' if the data is correct, click 'incorrect' if the data is incorrect.*' There were no practice trials, and participants then saw 96 passport frames, one at a time, and made their decision. Half the frames showed valid data and half invalid data. Each invalid data type (gender, year of birth, place of birth) was displayed in 16 passport frames, and no passport contained multiple errors. Faces were counterbalanced across the experiment such that each face occurred equally often in each condition. The task was self-paced, and typically took about 20 minutes to complete.

Results

Participants correctly classified 86% of the valid passports and 74% of the invalid passports; a significant effect of validity ($t(35) = 4.22$, $p<.001$). Figure 4 shows mean accuracy for each of the three invalid data types.

FIGURE 4 HERE PLEASE

A single factor ANOVA revealed a significant effect of data type ($F(2,70) = 38.19$, $p <.001$, $\eta_{\rho}^2 = .52$). Tukey HSD revealed significant differences between all pairs of conditions ($p < .05$ in all cases). In sum, viewers were best as spotting incorrect gender, intermediate at spotting incorrect place of birth, and worst at spotting incorrect year of birth.

Although this was not a speeded judgement task, we nevertheless recorded time taken to make these decisions. Time to verify a valid passport was significantly longer than time to spot an invalid one (means 9.7 sec vs 7.1 sec; SDs 5 and 2.6 respectively; $t(35) = 5.08$, $p<.001$). This is consistent with the fact that valid passports required exhaustive search of all three possible errors, while invalid passports could be classified correctly after any single dimension was identified as fraudulent. For the different types of fraudulent information, RTs followed accuracy data, with no sign of a speed/accuracy trade-off. Mean times to spot errors in gender, POB and YOB were 4.6 sec, 6.6 sec and 10.2 sec respectively ($F(2,70) = 43.30$, $p <.001$, $\eta_{\rho}^2 = .55$). Tukey HSD tests once again showed reliable differences between all conditions.

Discussion

Overall, viewers' accuracy on checking these passport documents was low – our participants spotted only 73% of fraudulent documents. Although performance was low in this check of passport internal consistency, it was nevertheless higher than in Experiment 1, where a second face was present – suggesting once again that identity checks using cards with photos *and* other data may be even harder in the field than laboratory studies suggest.

The accuracy of data checking for different types of information was straightforward. Gender errors were spotted very accurately, place of birth errors less so, and year of birth errors were the least easily detected. The high accuracy with sex judgements is consistent with previous research (e.g. Reddy, Wilken & Koch, 2004). Detection of year of birth errors is perhaps poorer than one might expect from research showing the general reliability of age estimations. However, we should note that this is not an age matching task. Passports show year of birth (Figure 1), and so computation of the bearer's age requires a calculation. This makes checking the appropriate age of the photo a several-step process - presumably accounting for the poor accuracy rate. Finally, the place of birth errors are interesting, because they do not require any comparison with the photo. Instead, simple checking for correct orthography of UK place names is shown to be rather error-prone in this situation.

Experiment 3

Introduction

Having established base rates for data checking on passport documents, we now return to the issue of how face matching and data checking interact in real documents. In this experiment we again asked participants to make both face matching and data checking judgements, as one might do when checking a person against photo-ID. We manipulated the validity of data shown in passport documents in the same systematic manner as Experiment 2. If, as suggested by Experiment 1, the presence of a face makes data checking harder, then we are particularly interested to know whether this is a general impairment, perhaps caused by attentional demands. Experiment 2 shows clear differences in the difficulty of detecting different types of fraudulent data – and general interference from a second face would impair performance across all conditions. On the other hand, it is possible that interference from a second face will only impair particularly difficult tasks, like age estimation, which rely on information within the face as well as the data section of the passport.

Method

Participants

Participants were 96 members of the Qualtrics panel (48 male), who were over 18, reported normal or corrected to normal vision and no dyslexia (mean age = 40.8, range = 18-73). To ensure their data checking was not compromised by a lack of familiarity with place names, participants were also required to be British Citizens who had lived in the UK for at least the last 10 years. Participants were reimbursed a small fee for their time.

Design and Procedure

As in experiment 1 this experiment was conducted online. Face stimuli were the same as those used in experiment 2: 96 face pairs from the GFMT long form, half male and half female pairs, with half matching and half mismatching. Participants saw 48 face-only pairs and 48 pairs in which one of the faces was embedded in a passport frame (see Figure 1). 24 of the passport frames contained valid data. Invalid data frames comprised 8 with the wrong gender (forename and M/F label), 8 with misspelled POB, and 8 with YOB 20 years too old. Faces were counterbalanced across the experiment such that each face appeared equally often in each condition.

Prior to the experiment participants were given two practice trials, a face-only pair, and a passport frame pair. They were asked to practise same/different face responses by selecting the appropriate button on screen. For the passport item, they were also asked to judge whether the personal information was correct with respect to the person shown in the passport. Instructions for this decision were as follows: '*If the data matches the image on the passport, e.g. correct gender, first name and year of birth click 'Correct'. If the data is factually incorrect or appears to contradict the image, click 'Incorrect''*.

Order of question (face match and personal data check) was fixed in this experiment (NB order had no effect in experiment 1), but order of stimuli was independently randomised in 12 versions of the task which were distributed equally across participants. Stimuli remained on the screen until responses had been made – i.e. until after both responses for the pairs including a passport frame. The task was self-paced and typically took about 30 minutes to complete.

FIGURE 5 HERE PLEASE

Results

Face Matching

Mean face matching accuracy is shown in Figure 5, with data broken down into match and mismatch trials. A 3 (presentation type: faces only, passport-valid, passport-invalid) x 2 (match type: same/different) within subjects ANOVA showed no main effect of presentation type ($F(2,190) = .51$, $p > .05$, $\eta_\rho^2 = .01$). However there was a main effect of match type

($F(1, 95) = 8.46$, $p <.005$, $\eta_\rho^2 = .08$) which was qualified by a significant interaction ($F(2,190) = 15.20$, $p <.001$, $\eta_\rho^2 = .14$).

Simple main effects analyses showed that presentation type affected same and different match items ($F(2, 380) = 6.13$, $p <.005$, $\eta_\rho^2 = .03$, and ($F(2, 380) = 11.36$, $p <.001$, $\eta_\rho^2 = .06$, respectively). As in Experiment 1, the presence of a passport frame leads to better performance for 'same' items and worse performance for 'different' items. Again, this pattern of data suggests that the presence of a passport frame biases viewers to respond 'same' to face pairs. This was confirmed with Signal Detection measures. d' was larger for the face only condition than for trials including valid or invalid passports (d' = 2.61, 2.32, 2.31 respectively, $F(2,95) = 12.63$, $p < .001$). Furthermore, criterion was very close to zero for face-only pairs, but significantly below zero for those embedded in passport frames (faces only, valid and invalid passport frames: criterion = 0.0, -0.18, -0.17 respectively, $F(2,95) = 12.52$, $p < .001$).

FIGURE 6 HERE PLEASE

Figure 6 shows mean face matching accuracy for the different types of invalid passport information. A 3 (invalid data type: gender/POB/YOB) x 2 (match type: same/different) ANOVA showed no main effect of data type ($F(2,190) = .68$, $p >.05$, $\eta_\rho^2 = .01$) or interaction ($F(2,190) = .45$, $p >.05$, $\eta_\rho^2 = .00$). Only the main effect of match type was significant ($F(1, 95) = 13.28$ $p <.001$, $\eta_\rho^2 = .12$), confirming higher overall performance for same-face trials – which is also clear in Figure 5.

Although this was not a speeded judgement, we also measured response times and observed that viewers spent longer looking at the face-only pairs than the pairs embedded in valid or invalid passport conditions (means 7.5 sec, 5.8 sec and 6.6 sec respectively; ($F(2,190) = 12.00$, $p <.001$, $\eta_\rho^2 = .11$). For faces-only and for valid-passport stimuli there was no difference in response time for same and different face pairs. However, when viewers were inspecting pairs embedded in invalid passport frames, they spend longer on different pairs than same pairs (means 7.6 sec and 5.6 sec respectively, ($F(1,285) = 17.72$, $p <.001$, $\eta_\rho^2 = .06$).

Data Checking

FIGURE 7 HERE PLEASE

Mean data checking accuracy scores are shown in Figure 7, broken down into trials in which the face pairs showed the same or different identities. A 2 (data type: valid/invalid) x 2 (match type: same/different) within subjects ANOVA showed a significant main effect of data type ($F(1,95) = 221.29$, $p < .001$, $\eta_\rho^2 = .70$), with no effect of match type ($F(1,95) = .70$, $p > .05$, $\eta_\rho^2 = .01$). These effects were qualified by a significant interaction ($F(1,95) = 23.63$, $p < .001$, $\eta_\rho^2 = .20$). Simple main effects analyses showed that presentation type affected both same and different match items ($F(1, 190) = 216.91$, $p < .001$, $\eta_\rho^2 = .53$, and ($F(1, 190) = 76.74$, $p < .001$, $\eta_\rho^2 = .29$), respectively.

FIGURE 8 HERE PLEASE

Most strikingly, these results show that detection of invalid passport information is very poor – and consistent with Experiment 1, the presence of a second (task irrelevant) face, reduces this performance to a level considerably lower than in a simple passport consistency check (Experiment 2). To explore this further, Figure 8 shows performance across the three different types of invalid data. A 3 (invalid personal data type: gender/POB/YOB) x2 (face match type: same/different) ANOVA showed a main effect of invalid personal data type ($F(2,190) = 124.57$, $p < .001$, $\eta_\rho^2 = .57$) and a main effect of face-match type ($F(1, 95) = 18.95$ $p < .001$, $\eta_\rho^2 = .17$) but no interaction ($F(2,190) = .57$, $p > .05$, $\eta_\rho^2 = .01$). As in Experiment 2, gender was the easiest false information to detect, but in this case POB and YOB were equally poor. Most interestingly, the presence of a matching face – which is task-irrelevant here – appears to reduce accuracy on a data-check. There appears to be a consistent tendency for 'same face' trials to produce a bias towards judging the personal data to be correct.

To examine this apparent bias, we compared signal detection measures for personal data-checking from same- and different-face pairs. We found no difference in d' (.98 vs .92, respectively, $t(95) = 1.36$, n.s.), but a significant difference in criterion (same-face and different-face trials, $c = -.86$ and $-.52$ respectively, $t(95) = 4.72$, $p < .001$), supporting the hypothesis that viewers are increasingly biased to judge personal information correct if this judgement is taken in the presence of matching face pairs. .

We also measured response times for data checking – though this was not a speeded judgment task, and there was no time pressure to make decisions. Time to verify a valid passport was marginally longer than time to spot an invalid one, though this difference did not reach significance (means 9.7 sec and 9.3 sec respectively ($F(1,95) = 0.84$, $p > .05$, $\eta_\rho^2 = .01$). As in Experiment 2, detection of incorrect sex was faster than incorrect POB, which was in turn faster than incorrect YOB (means 8.3 sec, 9.1 sec and 10.6 sec respectively, ($F(2, 190) = 9.85$ $p < .001$, $\eta_\rho^2 = .09$).

Discussion

This experiment shows clear effects, consistent with the earlier studies. As in Experiment 1 participants were compromised in performing a face match when a frame was added. Viewers were significantly more likely to identify same-face pairs as same, but less likely to identify mismatch trials, i.e. they are less likely to spot a fraudulent use. This is consistent across valid and all different types of invalid passport – so the simple fact of having a frame present induces this shift in face matching performance

Data matching performance is largely consistent with the earlier experiments. In particular, viewers are very poor at spotting invalid passports. Incorrect gender was spotted relatively accurately overall (about 78% of the time) but incorrect place of birth and year of birth was detected very infrequently (roughly 25% of the time) despite the fact that viewers had been alerted to look out for these problems. Interestingly, accuracy of data checking was affected by whether two faces showed same or different people – despite the fact that this was task irrelevant. Performance on confirming valid data was better when faces matched than when they did not. However, performance in spotting invalid data was better when the faces did not match. This suggests that the information from the faces ‘leaks into’ the data decision – which tends to be pulled in the direction of the faces. This is possibly because the faces are the more salient cue, even though their matching/non-matching status is independent of the data checking task. Interestingly, this effect of faces on data checking is consistent across the different types of invalid passport – i.e. it is the same in the easy gender check task as in the harder POB and YOB tasks. This suggests a strong effect of the secondary face, an effect which is reinforced by the fact that data checking in experiment 2, where there is no second face, results in fewer errors than observed here (though remains far from perfect).

General Discussion

The results presented here show, across multiple experiments, that face matching is affected by the presence of an ID context. The growing experimental literature on facial matching very often uses isolated face stimuli – for the good reason that researchers wish to study this process in the absence of potentially interfering material. However, in real world identity checks faces are very often compared to documents carrying other biographical information – as in passports, driving licences or workplace ID. The evidence presented here suggests that this significantly alters the patterns of accuracy obtained. When a face is embedded in a document, this seems to bias the viewer to make a 'same person' decision. Interestingly, this pattern seems to be due simply to the presence of the surrounding ID – the effect is consistent across documents containing valid information or any type of invalid information. This is clearly an important effect for practitioners, because it suggests that error detection for fraudulent ID-bearers may be under-estimated from currently available research.

The second important observation here is that the presence of a face match (two images) affects checking of biographical data on photo ID. Overall base-rates for checking the validity of biographical information are low (Experiment 2) but made worse when the check is required in the presence of a second face (Experiments 1 and 3). It is particularly interesting to note that the validity of the face match affects data-checking even when viewers were explicitly instructed to ignore the second face, as in Experiment 3. It seems that same-face pairs (i.e. those which are valid for the face) influence the likelihood that viewers will detect fraudulent biographical data. In short, if the faces match, viewers are more likely to say that the data is correct too – even when it is not. Interestingly, this effect is completely consistent across different types of data-invalidity, and is independent of whether the biographical-error is generally easy to spot (gender) or hard to spot (place and year of birth).

Could the effects of faces on data checking be due to dual-task interference (e.g. Pashler, 1994)? At first glance, this looks rather unlikely. Although the face matching and data-checking tasks shared stimuli they were always presented separately in these experiments, and so there seems no opportunity for one task to be effected by load due to the other.

However, it is possible that the presence of a face leads to automatic processing (Farah, Wilson, Drain & Tanaka, 1995). If so, there could then be competition for resources to process face and biographical-data stimuli. The traditional view is that the shallower of two tasks is affected by dual-task interference (Jones, Miles & Page, 1990), and so an attention-demanding face task (albeit an incidental one) may affect performance on other tasks. Some research suggests that faces are processed mandatorily at a semantic level, and not just for superficial visual characteristics (Boehm, Klostermann, Sommer & Paller, 2006; Burton, Kelly & Bruce, 1998). This may provide a route to understanding how an apparently superficial face-matching process could interfere with apparently complex tasks such as calculating someone's age and making a judgement about its veracity.

These results have some implications for practitioners. Of course, these studies are performed on non-expert viewers, and it will be important to establish in future whether these effects generalise to people who conduct ID checks professionally (White, Phillips, Hahn, Hill & O'Toole, 2015). Within the fingerprint matching context, individuals without expertise have been shown to make false match decisions based on the gender or ethnicity of the 'accused' (Smalarz, Madon, Yang, Guyll, & Buck, 2016). Stereotypical beliefs about crimes and the types of people who commit them 'leak through' into the fingerprint matching judgement despite having no bearing on the decision. Research carried out with fingerprint experts has shown that they are also vulnerable to contextual biases such as prior match or mismatch decisions (Dror, Charlton & Péron, 2006). This would suggest that experts in face matching and ID checking may also be vulnerable to these biases. For such people, it may be possible to separate data checks and face checks in the workflow, and our results suggest this should be done if possible. The results also have implications for the relationship between laboratory experiments and real world settings, providing additional support to the concerns raised by researchers in forensic science. Our data emphasise that if researchers intend to generalise their results outside the lab, then it is important to incorporate all relevant task demands. Apparently unrelated components of the task can have significant effects on each other, and we have demonstrated here that previous research on face matching may have under-estimated the problem for those performing the task professionally.

References

Anastasi, J. S., & Rhodes, M. G. (2006). Evidence for an own-age bias in face recognition. *North American Journal of Psychology, 8*, 237–252.

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*(5), 625-627.

Boehm, S. G., Klostermann, E. C., Sommer, W., & Paller, K. A. (2006). Dissociating perceptual and representation-based contributions to priming of face recognition. *Consciousness and cognition*, *15*(1), 163-174.

Bruce, V., Ellis, H. D., Gibling, F., & Young, A. (1987). Parallel processing of the sex and familiarity of faces. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *41*(4), 510-520.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339-360.

Burton, A. M., Kelly, S. W., & Bruce, V. (1998). Cross-domain repetition priming in person recognition. *Quarterly Journal of Experimental Psychology*, *51A*(3), 515–529.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286-291.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482–505.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*(1), 74-78.

Farah, M. J., Wilson, K. D., Drain, H. M., & Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision research*, *35*(14), 2089-2093.

Fender, M. (2008). Spelling Knowledge and Reading Development: Insights from Arab ESL Learners. *Reading in a Foreign Language*, *20*(1), 19-42.

Heyer, R., & Semmler, C. (2013). Forensic confirmation bias: The case of facial image comparison. *Journal of Applied Research in Memory and Cognition*, *2*(1), 68-70.

HM Passsport Office (2014). Her Majesty's Passport Office Annual Report and Accounts 2013-2014. Retrieved from:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/332680/HMPO_ARA_2013-14_v5_1_FINAL_WEB.pdf

Jenkins, R., Lavie, N., & Driver, J. (2003). Ignoring famous faces: Category-specific dilution of distractor interference. *Perception & Psychophysics*, *65*(2), 298–309.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577-596.

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology*, *4*(2), 89-108.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*(1), 42-52.

Kemp, R., Towell, N., & Pike, G. (1997). When Seeing should not be Believing : Photographs , Credit Cards and Fraud. *Applied Cognitive Psychology*, *11*(3), 211–222.

Lavie, N., Ro, T., & Russell, C. (2003). The role of perceptual load in processing distractor faces. *Psychological science*, *14*(5), 510-515.

Leslie, L., & Thimke, B. (1986). The use of orthographic knowledge in beginning reading. *Journal of Literacy Research*, *18*(3), 229-241.

Megreya, A. M., & Burton, A.M. (2006). Unfamiliar faces are not faces: evidence from a matching task. *Memory & Cognition*, *34*(4), 865–76.

Megreya, A. M., and Burton, A.M. (2008). Matching faces to photographs: poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied 14*(4), 364-372.

Menneer, T., Stroud, M. J., Cave, K. R., Li, X., Godwin, H. J., Liversedge, S. P., & Donnelly, N. (2012). Search for two categories of target produces fewer fixations to target-color items. *Journal of Experimental Psychology: Applied, 18(4)*, 404-418.

Moyse, E. (2014). Age Estimation from Faces and Voices: A Review. *Psychologica Belgica*, *54*(3), 255-265

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition*, *26*(1), 146-160.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, *116*(2), 220-244.

Reddy, L., Reddy, L., & Koch, C. (2006). Face identification in the near-absence of focal attention. *Vision Research*, *46*(15), 2336-2343.

Reddy, L., Wilken, P., & Koch, C. (2004). Face-gender discrimination is possible in the near-absence of attention. *Journal of Vision*, *4*(2), 4.

Smalarz, L., Madon, S., Yang, Y., Guyll, M., & Buck, S. (2016). The Perfect Match: Do Criminal Stereotypes Bias Forensic Evidence Analysis? *Law and Human Behavior,* 40(4), 420-429.

Voelkle, M. C., Ebner, N. C., Lindenberger, U., & Riediger, M. (2012). Let me guess how old you are: Effects of age, gender, and facial expression on perceptions of age. *Psychology and Aging, 25*, 265– 277.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport Officers' Errors in Face Matching. *PloS one*, *9*(8), 1-8.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*(20151291).

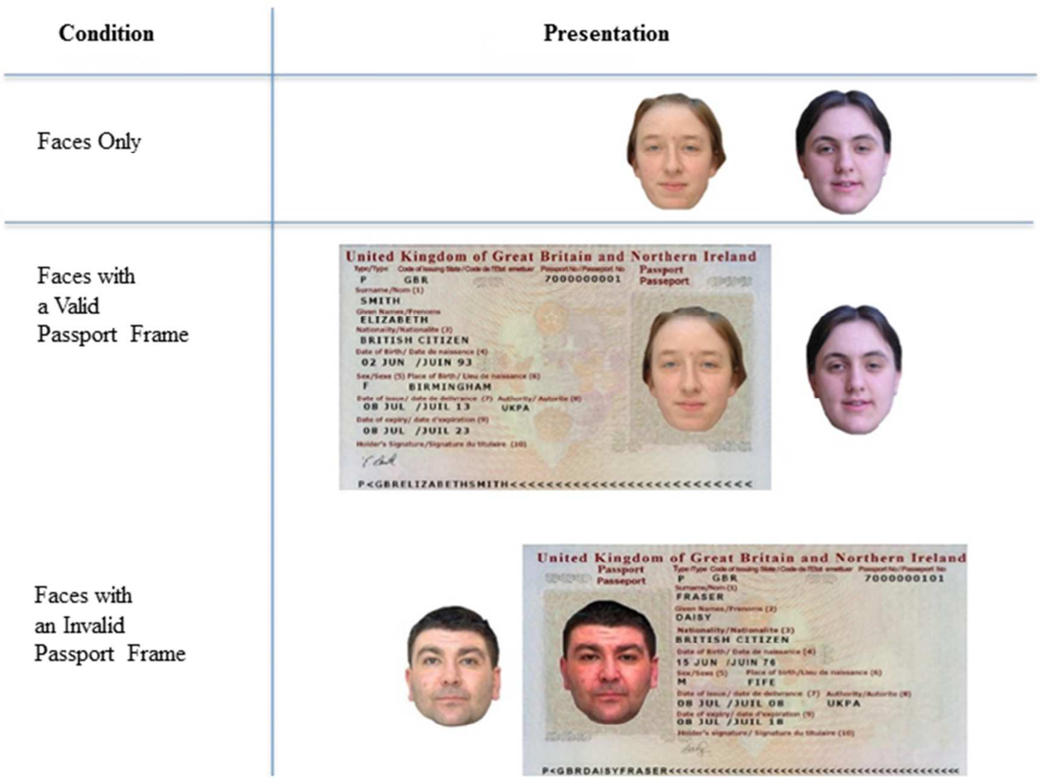

Figure 1: Items from GFMT in normal presentation, and with one face embedded in a passport frame. All three types of presentation were used in Experiments 1 and 3. For Experiment 2 (data checking) a single passport frame was used, with no second face.

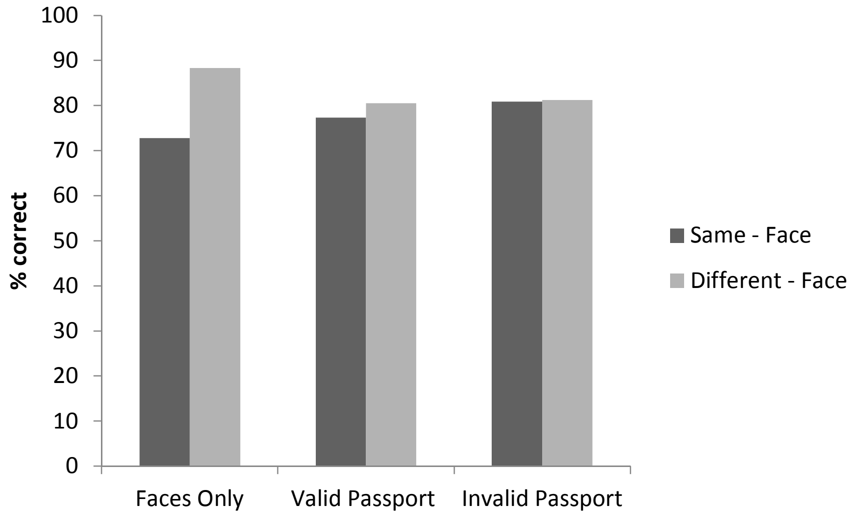

Figure 2. Mean face matching accuracy across conditions (Experiment 1).

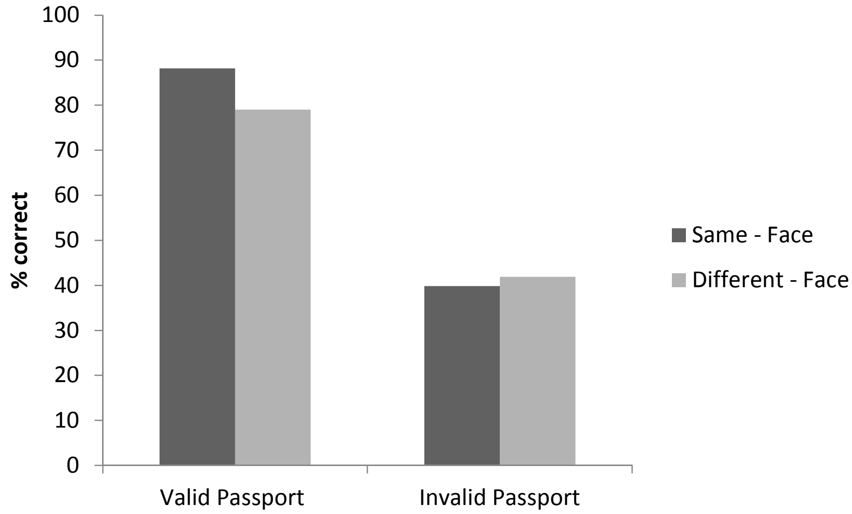

Figure 3. Mean accuracy for data checking (Experiment 1).

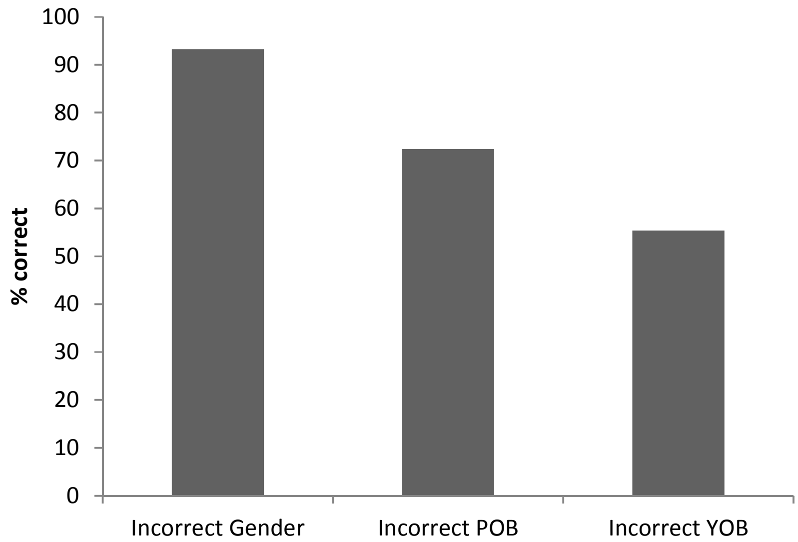

Figure 4. Mean data checking accuracy for invalid data types (Experiment 2).

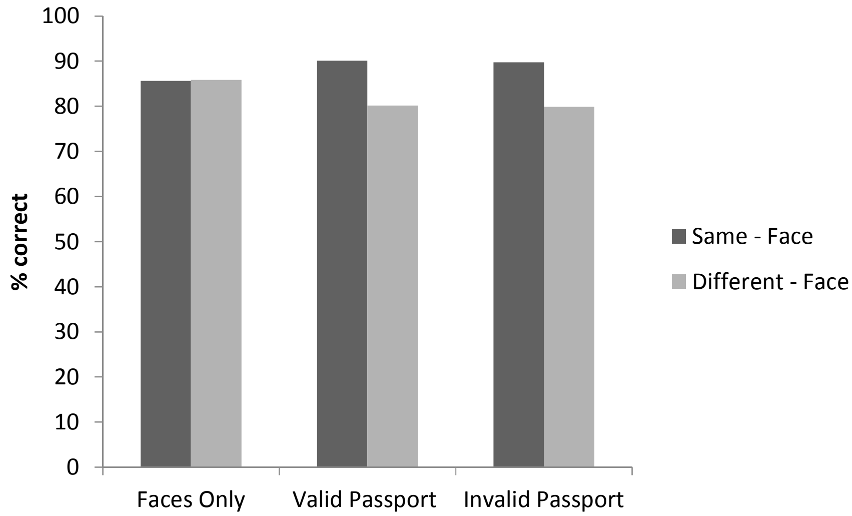

Figure 5. Mean face matching accuracy across conditions (Experiment 3).

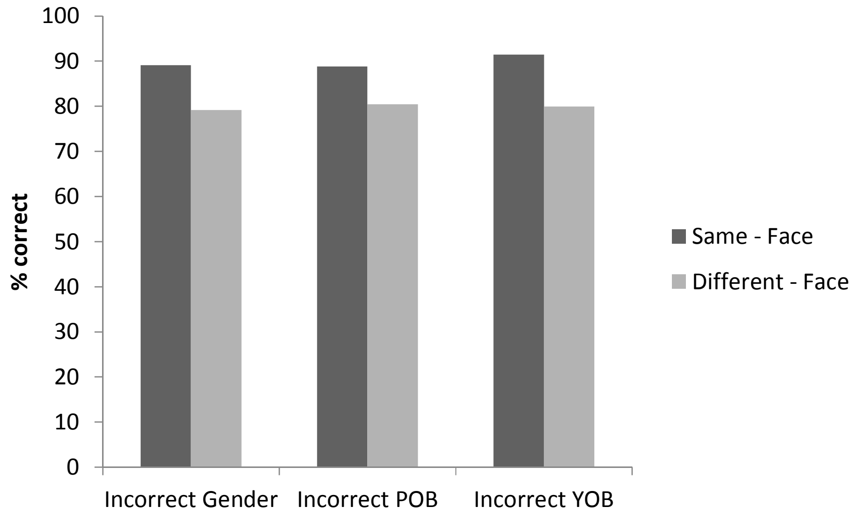

Figure 6. Mean face matching accuracy across types of invalid data (Experiment 3).

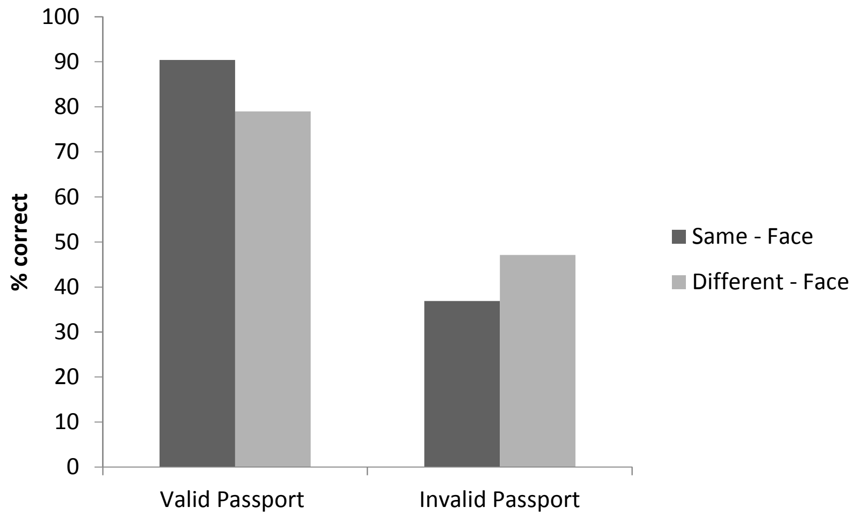

Figure 7. Mean accuracy for personal data checking (Experiment 3), in the presence of a task-irrelevant same-face or different-face pair.

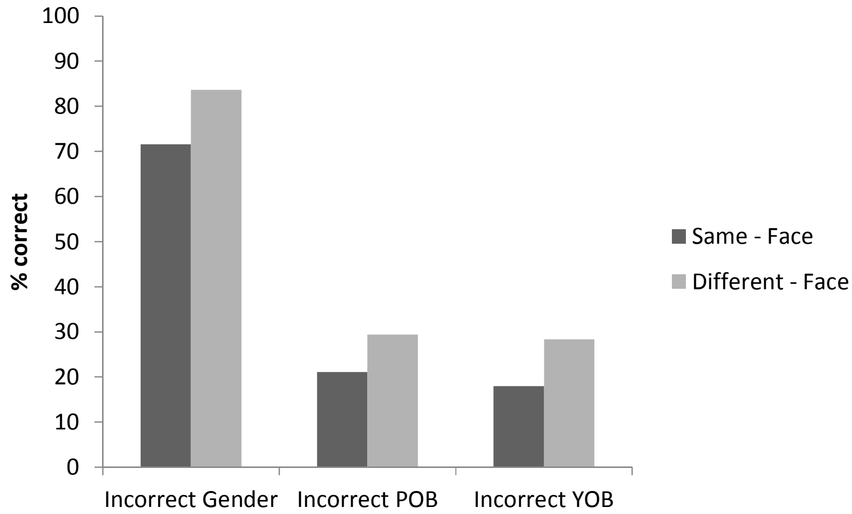

Figure 8. Mean data checking accuracy for invalid data types (Experiment 3), in the presence of a task-irrelevant same-face or different-face pair.