Vision Research 157 (2019) 159–168

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

# Other-race faces are given more weight than own-race faces when assessing the composition of crowds

Ian M. Thornton[a,*], Duangkamol Srismith[b], Matt Oxner[c], William G. Hayward[d]

[a] *Department of Cognitive Science, University of Malta, Malta*
[b] *Department of Psychology, Swansea University, UK*
[c] *School of Psychology, The University of Auckland, New Zealand*
[d] *Department of Psychology & The Australian Research Council Centre of Excellence in Cognition and Its Disorders, University of Hong Kong, Hong Kong*

ARTICLE INFO

*Keywords:*
Face perception
Race
Other-race effect
Other-race bias
Facial groups
Ensemble representations

ABSTRACT

In two experiments we examined the performance of Asian and Caucasian participants as they were asked to estimate the ethnic composition of arrays of 16 concurrently presented faces. Across trials we systematically varied the physical proportion of Asian and Caucasian faces presented in the arrays using the method of constant stimuli. The task was to explicitly indicate which group was in the majority. The position of the 16 faces within the array were continuously shuffled using a 4 × 4 moving grid to block explicit enumeration. Measures of bias and sensitivity were estimated by fitting cumulative normal distributions to individual response data. Consistent with recent findings on "ensemble" face processing, all participants were able to make group estimates quite accurately. This was true using both full-colour, non-normalised, headshots (Exp1) and centre-apertured, normalised, grey-scale images (Exp2). However, the main finding was that performance estimates from the two groups of participants did not overlap. Specifically, patterns of bias suggest that other-race faces are weighted more heavily than own-race faces (Exps 1 & 2), while sensitivity is better for groups instructed to decide if the other-race, rather than own-race, is more numerous (Exp 2). To our knowledge, these are the first demonstrations of other-race biases affecting decisions that have to be made about groups of faces.

## 1. Introduction

It is well-established that the processing of individual faces can be affected by the race of the observer. The other-race effect (ORE) refers to situations where perception and memory is better for faces of one's own race compared to other-race faces (Malpass & Kravitz, 1969). Conversely, an other-race search advantage (ORSA) has been reported, where faces from another race can be found more efficiently than own-race faces (Levin, 2000; Sun, Song, Bentin, Yang, & Zhao, 2013). A number of explanations have been proposed for these race-dependent effects, chief among them being variations of the "contact hypothesis" (Allport, 1954), suggesting that the quantity and quality of interactions with a racial group is the most important factor (Tanaka, 2013).

Recently, Laurence, Zhou, and Mondloch (2016) demonstrated that such race-dependent effects also extend to situations where judgements involve multiple images of faces. In their study, participants completed a perceptual sorting task in which 40 images of two unfamiliar individuals had to be placed into unique identity piles. Consistent with previous research, few participants correctly created two piles, demonstrating the difficulty of extracting consistent identity information given within-person variability in photographs (Jenkins, White, Van Montfort, & Burton, 2011). More importantly, across two experiments, the number of perceived identities was higher when sorting other-race faces than when sorting own-race faces. This finding suggests that there may be reduced perceptual/representational precision when resolving the identity of other-race faces across multiple instances.

In the current work we also use multiple faces to explore race-dependent effects. However, rather than explicitly asking participants to make identity judgements, we asked them to assess the racial composition of groups of faces that were shown concurrently. Our question was whether estimates of which racial group was in the majority would be influenced by the race of the participant.

This research direction was directly inspired by recent studies of "ensemble face" processing. When presented with a brief display containing multiple faces, observers can rapidly extract the average emotion, gender or identity present in the group (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007, 2009). Such ensemble processing has been shown to occur both when the multiple faces are distributed across space (Haberman & Whitney, 2007) or across time (Haberman, Harp, & Whitney, 2009). Jung, Bülthoff, and Armann (2017) recently

[*] Corresponding author.
*E-mail address:* ian.thornton@um.edu.mt (I.M. Thornton).

https://doi.org/10.1016/j.visres.2018.02.008
Received 15 September 2017; Received in revised form 15 February 2018; Accepted 16 February 2018
Available online 24 March 2018

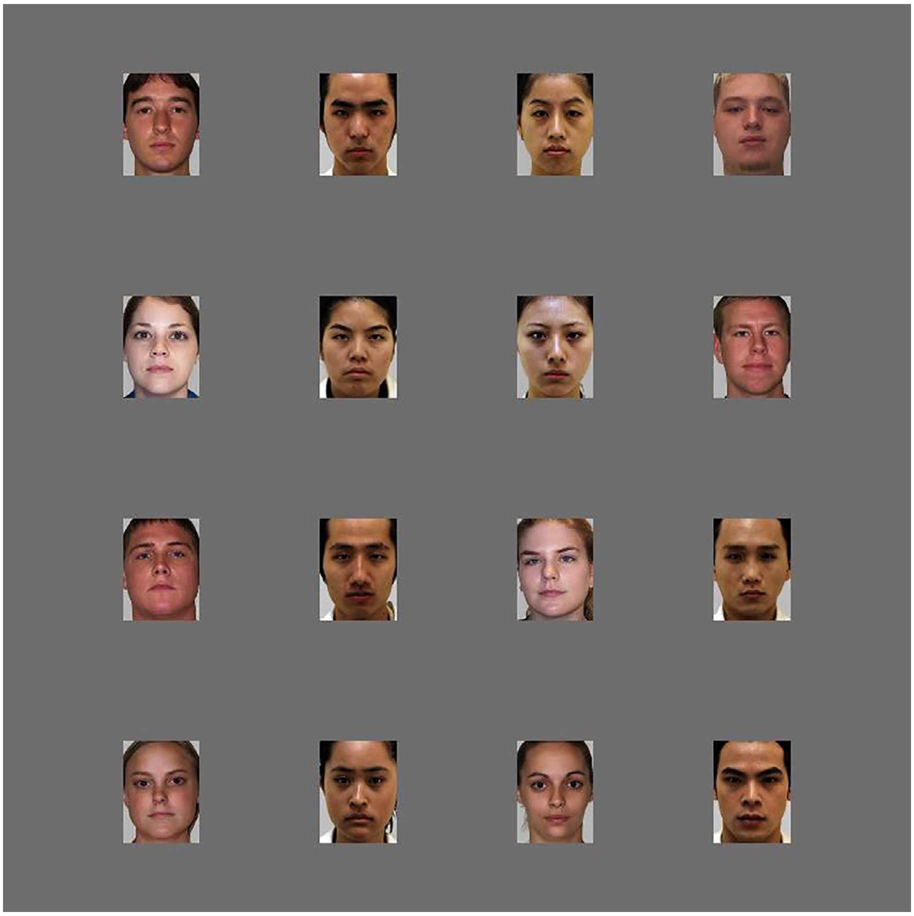

**Fig. 1.** Example stimulus array with equal Asian and Caucasian faces. On each trial, 16 individuals were selected from a database and the faces shuffled position every 250 ms to reduce the tendency to subsample the display. Gaps between each image were initial wide enough so that the outer faces were close to the edge of the screen and contracted within 4 s so that all sixteen faces were adjacent just before disappearing.

extended this line of research by showing that groups of Caucasian participants were able to estimate average race in sets of morphed heads taken from an Asian-Caucasian continuum. More generally, it is thought that ensemble representations provide an efficient means to reduce perceptual bandwidth by providing precise estimates of the mean of a set of items without the need to visually inspect every member (Alvarez, 2011; Ariely, 2001; Whitney & Yamanashi Leib, 2018).

In the two experiments reported here, we asked participants to make explicit judgements about the racial composition of dynamic arrays that varied in the proportion of Asian and Caucasian identities they contained. The main goal of the current study was not to directly probe the creation or nature of ensemble "race" representations, but rather to determine whether participant race influenced performance. We ran studies both in the United Kingdom (UK) and in Hong Kong (HK), and always included equal numbers of Asian and Caucasian participants from each location.

Initially, we needed to establish whether observers could estimate the racial composition of the multiply-presented faces in our particular displays. That is, we wanted to know whether participants were able to rapidly (i.e., without explicitly counting) and accurately judge whether more Asian or more Caucasian faces were contained in our arrays. This would not only serve to corroborate the finding of Jung et al. (2017) – that race is another type of information that can be rapidly extracted from groups of faces (i.e., in addition to emotion, gender, identity) – using very different stimuli and tasks, but would also make it possible to examine our second, and more central theme.

That is, we wanted to know whether estimates of the racial composition of a group of faces would be affected by the race of the participant. This question not only has the potential to increase our knowledge about race-dependent effects, i.e., confirming that they occur at the group level (Laurence et al., 2016), but may also shed light on the mechanisms that underlie the extraction of summary representations, for example, are such representations influenced by individual differences (Haberman, Brady, & Alvarez, 2015)?. Additionally, we examined whether such estimates might be affected by recent exposure to the other race (an ex-pat effect, Experiment 1) or the removal of all low-level images cues, so that only the shape and configuration of the central facial features were available (Experiment 2).

## 2. Experiment 1

In our first experiment, we asked observers to explicitly indicate whether an array of 16 faces contained more Asian or more Caucasian faces. The array size was modelled on previous ensemble studies (e.g., Haberman & Whitney, 2009) and was designed to be large enough to minimize explicit enumeration without overwhelming the face processing system. By varying the physical proportion of the two racial categories across trials we could estimate individual points of subject equality (PSEs), the point at which a participant perceived the arrays to have equal numbers of Asian and Caucasian faces. This was our main dependent measure, although we also measured the overall precision and speed of responses.

As our interest was in the ability to "rapidly" assess the racial composition of a crowd, we needed to avoid the possibility that participants could slowly count through the arrays. In contrast to typical ensemble studies, where face arrays are usually shown very briefly, for example between 50 and 2000 ms (Haberman & Whitney, 2007, 2009), we presented our arrays for up to 4 s. Rather than using duration to reduce the possibility of explicit enumeration or display sub-sampling, we used two "dynamic" manipulations. First, the 16 constant identities selected on a given trial were spatially shuffled every 250 ms within the array. Second, the whole display contracted, with the gaps between images reducing from the edge of the screen until faces were adjacent, at which time the screen was blanked, after approximately 4 s.

The idea behind both of these forms of animation was to allow the visual system ample time to form an estimate, given a constant racial proportion, while still minimizing explicit strategies. We discuss the possible consequences of using such animation in the General Discussion.

As can be seen in Fig. 1, in this initial experiment we did not crop or normalise the facial images in any way. That is, each facial image was displayed in full colour and showed both internal (e.g., eyes, nose, mouth and their configuration) and external (e.g., ears, head shape) features as well as hair. Estimates of the racial composition of an array could thus be based on a number of factors, including relatively low-level differences, such as skin tone and hair colour. In Experiment 2, we explicitly address this aspect of the design.

Finally, we ran the study both in the United Kingdom (UK) and in Hong Kong (HK) and in each centre we had equal numbers of Asian and Caucasian observers. Thus, we could independently assess the effects of race and exposure/expertise.

## 3. Methods

### 3.1. Ethics statement

All aspects of the experimental protocols were reviewed and approved by the University of Swansea Psychology Department ethics committee and the study was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

### 3.2. Participants

A total of 40 (20 female) participants took part in the study. Ten Asian (Chinese) and 10 Caucasian (White European/North American) individuals were run in the UK and an identical group were run in HK. All observers had normal or corrected to normal vision, gave written informed consent and were naïve as to the purpose of the study.

### 3.3. Equipment

In both locations, stimuli were presented on standard LCD monitors with effective refresh rates of 60 Hz. Stimuli presentation, response collection and data recording were controlled by custom written software developed in MATLAB, using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Participants sat approximately 120 cm away from the screen and made responses via a standard USB keyboard.

### 3.4. Stimuli

Stimuli consisted of arrays of full-colour images of faces taken from two publically available databases. Caucasian faces were taken from the Center for Vital Longevity Face Database (Minear & Park, 2004) and Asian faces were chosen from the Taiwanese Facial Expression Image Database (Chen & Yen, 2007). In Experiment 1, images were not cropped or normalised in any way. Each facial image subtended approximately 0.8° × 1.2° visual angle.

As shown in Fig. 1, the experimental display consisted of an array of 16 images that could vary in the proportion of Asian and Caucasian faces. From each database, 20 male and 20 female faces (80 faces in total) were used as the pool from which 16 faces were randomly selected on each trial. The only constraint, aside from the parametrically varied racial proportion, was that each array always contained the same number of male and female faces. At its widest separation, the whole array subtended approximately 10°, contacting to approximately 4° within a 4 s viewing period. The position of each face within the grid was randomly reassigned every 250 ms.

### 3.5. Task

On each trial, the participants had to indicate, using an assigned key, whether there were more Asian or more Caucasian faces present in the display. Key assignment was counterbalanced across participants. Accuracy was stressed over speed, and participants were instructed to take as much time as required to make their decision. Note that the display was only visible for approximately 4 s, but responses were still recorded past this limit. Less than 8% of all responses occurred when the stimuli were not visible. The next trial began 500 ms from the time of response.

The actual proportion of races within a given trial varied in steps of 2 (1 male and 1 female) from 0 Asian/16 Caucasian to 16 Asian/0 Caucasian. Each of the nine display proportions was presented 30 times for a total of 270 trials. Trial order was randomized separately for each participant.

### 3.6. Analysis

Cumulative normal distributions were fitted to the data of individual participants using the Palamedes Toolbox (Prins & Kingdom, 2009) for MATLAB. Goodness of fit was determined via the bootstrapping method described by Wichmann and Hill (2001). Only 4 out of 40 curves fell outside of the 95% confidence intervals, suggesting very good approximations. Analysis was performed with and without the poorly fitting data, and as the results were qualitatively identical, the full dataset are reported below.

We extracted two parameters from the fitted curves: the point of subjective equality (PSE) and the just noticeable difference (JND). The former provides an estimate of bias away from the physical mid-point (veridical) array containing 8 Asian and 8 Caucasian faces, and the latter measures the precision with which such responses are made. For the JND measures, larger values indicate less precision. For both parameters the basic unit of measurement is the number of face images on a scale of 0 to 16. These parameters were analysed in separate 2 (Location: HK, UK) × 2 (Participant Race: Asian, Caucasian) between-subjects ANOVAs. Although we did not emphasize speed of response during the task, as the displays dynamically changed during the course of a trial, we analysed reaction time using the same 2 × 2 model to explore whether there were systematic differences in exposure time that might influence the other measures.

## 4. Results

Fig. 2A presents a summary of the response data for all participants, collapsed across Location. Data have been plotted, arbitrarily, as a function of the number of Asian faces. The two curves represent Asian and Caucasian participants, respectively. It is clear from the figure that the response distributions for the two groups of participants do not completely overlap. Fig. 2B summarizes the PSE values extracted from individual curves, plotted as a function of both Location and Participant Race. The 2 × 2 ANOVA on these data showed a main effect of Participant Race, F(1,36) = 9.2, MSE = 1.2, $p < .01$, pEta = 0.20, indicating that Asian participants (M = 9.0, SEM = 0.21) required consistently more Asian faces in the array than Caucasian participants (M = 7.9, SEM = 0.21) in order to perceive subjective equality. There was no main effect of Location, F(1,36) = 0.7, MSE = 1.2, n.s., and no interaction, F(1,36) = 0.1, MSE = 1.2, n.s.

For the JND analysis (Fig. 2C) there was a marginal main effect of Participant Race, F(1,36) = 3.8, MSE = 1.2, p = .06, pEta = 0.10, with Caucasian participants (M = 2.4, SEM = 0.24) being slightly more sensitive to the racial composition of the arrays than Asian participants (M = 3.1, SEM = 0.24). Again, there was no main effect of Location, F (1,36) = 0.9, MSE = 1.2, n.s., and no interaction, F(1,36) = 0.2,

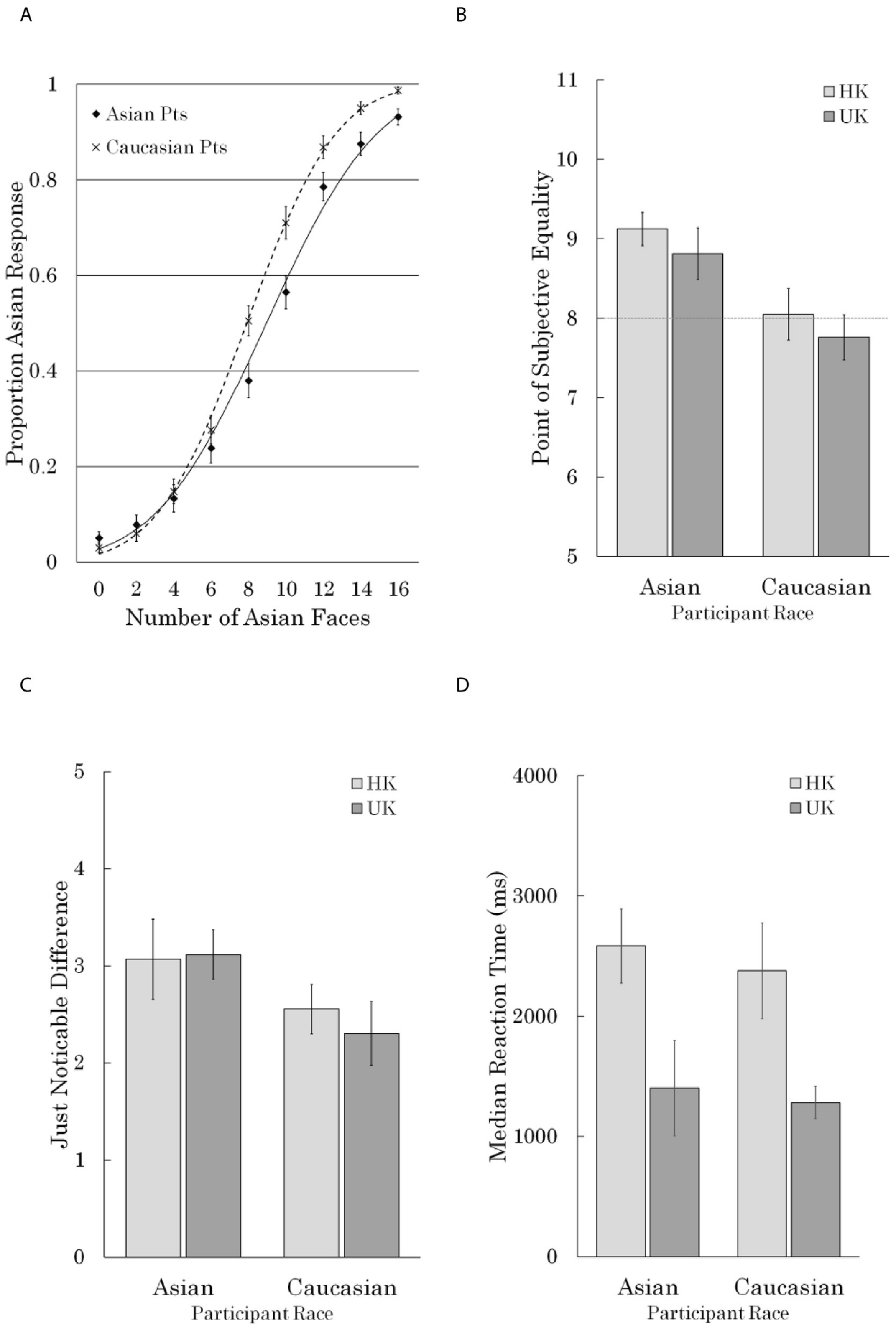

**Fig. 2.** Results of Experiment 1. A) Proportion Asian Responses collapsed across Location; B) PSE as a function of Location and Participant Race. The dotted line shows veridical equality; C) JND as a function of Location and Participant Race; D) Median RT as a function of Location and Participant Race. Error bars are standard error of the mean.

MSE = 1.2, n.s.

The median RT data are shown in Fig. 2D. Here, there was a main effect of Location, with participants tested in the UK (M = 1342 ms, SE = 103 ms) responding more than a second faster than those in HK (M = 2482 ms, SE = 246 ms), F(1,36) = 17.5, MSE = 742077, p < .01, pEta = 0.33. There was no main effect of Participant Race, F (1,36) = 0.4, MSE = 742077, n.s., and no interaction, F(1,36) = 0.1, MSE = 742077, n.s.

## 5. Discussion

As outlined in the Introduction, the goal of this research project was to answer a number of basic questions about the ability to estimate race from groups of faces. The results of Experiment 1 provide initial answers to three of these questions. First, the general pattern of results shown in Fig. 2A indicates that people were able to accurately estimate the majority race presented in these brief displays. Performance is close to ceiling at the extreme ends of the distributions (i.e. when arrays contained 100% Asian or Caucasian faces) and shifts systematically as the racial proportion changes.

Second, there was a consistent effect of Participant Race on performance. This means that when looking at the same physical displays, Asian and Caucasian participants made different decisions about the racial composition of the crowds of faces. For Caucasian participants, the average PSE was very close to veridical. Note that the veridical point of equality was a display containing 8 Asian and 8 Caucasian faces. For the Asian participants, this estimate was shifted significantly to the right. This means that Asian participants appeared to give more weight to Caucasian faces, requiring more Asian faces to be present in the display to perceive equality.

Sun et al. (2013) found a similar asymmetrical pattern of results in a visual search paradigm. They measured the efficiency of search for own and other-race faces and only found a search asymmetry for Chinese participants, an effect similarly driven by their sensitivity to other-race faces. It is unclear what drove this asymmetry in the search data or why we should find a similar pattern with the current task. Of course, some caution should be urged in over-interpreting "veridical" performance in our task, as the physical mid-point (8 Asian/8 Caucasian) may not necessarily equate to an unbiased perceptual mid-point. That is, performance by participants from a third race (e.g., African) might fall in the middle (i.e. 8.5) of our two current groups.

More generally, biases as measured in the current task could originate at the perceptual level (e.g., differential sensitivity to facial cues), at the decision/response level (e.g., how to respond when uncertain) or some combination of both. It may seem more reasonable to assume that if only one group of participants has a clear bias, that its origin is at the response/decision level, rather than a basic, perceptual level, unique to one race. In the next experiment we try to explicitly tease apart these two influences by modifying the nature of the task.

The third question concerned the impact of recent exposure. In Experiment 1, we ran our first twenty participants in the UK and had initially attributed the asymmetrical shift in responses for Asian participants to some form of "ex-pat" effect. Our thought was that recent exposure to other-race faces when living in a new country may have increased their salience and given them more weight when trying to estimate the composition of our displays. However, when we ran the identical design in HK, rather than finding a mirror reversal (as now Caucasian participants were the ex-pats) the patterns across race were identical to those seen in the UK sample. As there were no Location × Participant Race interactions for any of our dependent measures, there appears to be little role of recent exposure with the current task.

In addition to the above patterns of results, one unexpected finding was the difference in RT between participants in the two locations. We have no clear explanation for this finding. Participants in both locations were given the same basic instructions about the task and method of response and were given equal amounts of practice with the task. One possibility is that that our participants in HK may have had more prior experience with taking part in face memory experiments. They may have tried to maximize their exposure to the facial identities in expectation of a later test of recognition memory. Alternatively, some subtle difference in the way the displays were initially demonstrated in the two locations may have set different expectations of when to respond. In any event, as there were no Location main effects and no Participant Race × Location interactions on our two other dependent measures this suggests that the current task is actually quite insensitive to absolute exposure duration.

## 6. Experiment 2

The main finding of Experiment 1 suggests that participants systematically estimate the racial composition of identical arrays differently, depending on their own racial background. To our knowledge, this is the first demonstration of such an other-race bias when making decisions about groups of faces. The first goal of Experiment 2 was to attempt to replicate this finding with a new set of stimuli and two new groups of participants.

As noted in the discussion of Experiment 1, the differences between groups in terms of PSE, and in particular, the apparently asymmetrical nature of the shift – such that Caucasian participants were close to veridical and Asian participants had an other-race bias – made us question whether the source of such a difference lay at the perceptual level or the decision making/response level. For example, the observed shift could have arisen if Asian participants chose to systematically endorse the "other race" whenever they were uncertain but Caucasian participants had no consistent strategy. As we had asked participants to explicitly make an "Asian" or "Caucasian" response, it is conceivable that Asian participants felt more "socially comfortable" endorsing the other-race when they were guessing.

To try and dissociate this sort of response bias from a perceptual bias – in which same and other race faces are given different weights by the visual system – in Experiment 2, we changed the nature of explicit decision that needed to be made. Specifically, half of the participants of each race were told that their target category were Asian faces and their distractor category were Caucasian faces. The other half were given the reverse mapping. For all participants, the task was now to judge if there were more target faces or distractor faces on each trial. Thus, we explicitly prioritized one race, controlling and counterbalancing how this was mapped. Importantly, participants were also explicitly told to endorse the target category whenever they were unsure of which response to make. If the asymmetrical shift in Experiment 1 has been caused only by a response bias under uncertainty, here we would expect an identical shift in favour of the target category in both groups of participants, irrespective of Participant Race. If PSE differences between Asian and Caucasian observers still persist, then this would more strongly hint at a perceptual locus.

Another possible limitation of Experiment 1 was our use of full-face colour photographs that had not been normalised. In Experiment 2, we wanted to more tightly control the cues to race that were available. As described in more detail below, we cropped and normalised the faces to remove all differences between the sets except for variation in the internal facial features and their configuration. This reduces the possibility that our two Participant Races were using different cues, and removes the influence of low-level image artefacts.

## 7. Methods

### 7.1. Ethics statement

All aspects of the experimental protocols were reviewed and approved by the University of Swansea Psychology Department ethics committee and the study was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

### 7.2. Participants

A total of 32 (16 female) participants took part in the study. As Participant Race and Location did not interact with any of the measures in Experiment 1, we did not cross these factors in Experiment 2. Sixteen Asian (Chinese) individuals were run in HK and 16 Caucasians (White European/North American) were run in the UK. All observers had normal or corrected to normal vision, gave written informed consent and were naïve as to the purpose of the study.

### 7.3. Equipment

All aspects of the experimental set-up were the same as in Experiment 1.

### 7.4. Stimuli

Forty Asian facial images (20 female/20 male) were obtained from the HKU Psychology Department stimulus bank and 40 Caucasian faces (20 female/20 male), were made available by the Face Perception Lab at Brock University, CA. (https://brocku.ca/psychology/research/faceperceptionlab.htm). In contrast to Experiment 1, we did not show full colour images, but modified the stimuli in several ways to ensure that the only cues to race came from internal features and their configuration. First, each image was cropped using an oval aperture to remove details of head shape and hair. All 80 images were then equated in terms of mean luminance and contrast using the default parameters of the lumMatch function from the SHINE Matlab Toolbox (Willenbockel et al., 2010). Finally, a Gaussian mask was used to blend each image into the uniform gray background. Each of the cropped images subtended approximately 0.64° × 0.95° visual angle, a reduction of approximately 20% compared to Experiment 1.

### 7.5. Task & analysis

Each group of participants was explicitly told which race was their target and which was the distractor race. On each trial, they had to indicate, using an assigned key, whether there were more Target or more Distractor faces present in the display. They were also explicitly told that when they were unsure of which decision to make, they should endorse their target category. Target/distractor key assignment was counterbalanced across participants. All other aspects of the task, design & analysis were the same as in Experiment 1, except that the statistical model now became a 2 (Target Race) × 2 (Participant Race) ANOVA.

## 8. Results

Fig. 4A shows the response data for all participants, collapsed across Target Race. As in Experiment 1, data have been plotted, arbitrarily, as a function of the number of Asian faces in each array. Compared to Fig. 2A, these average slopes look shallower – possibly reflecting increased difficulty with cropped and normalised faces – but again, there is considerable lack of overlap for much of the range, suggesting differences between the two groups of participants.

Fig. 4B shows that the pattern of PSEs as a function of Target Race and Participant Race. As predicted, instructing participants to endorse their target category when uncertain led to clear shifts in the PSE as a function of Target Race. Specifically, there was a main effect of Target Race, F(1,28) = 16.5, MSE = 1.9, $p < .001$, pEta = 0.4, such that participants with Asian Targets (M = 6.99, SE = 0.3) needed fewer Asian faces (compared to the veridical point of 8) to be physically present in order to perceive a balanced array. Conversely those with Caucasian targets (M = 8.97, SE = 0.3) needed more Asian faces and therefore fewer Caucasian faces.

Over and above this influence of Target Race, as in Experiment 1, there was still a main effect of Participant Race, F(1,28) = 6.6, MSE = 1.9, $p < .05$, pEta = 0.19. In contrast to Experiment 1, the influence of Participant Race was symmetrical, with both the Asian (M = 8.6, SE = 0.35) and the Caucasian (M = 7.35, SE = 0.35) participants shifting away from veridical (i.e. 8 faces), requiring fewer faces of the other race to perceive equality. As can be seen in Fig. 4B, across both levels of Target Race, Asian participants require fewer Caucasian faces, and Caucasian participants require fewer Asian faces. There was no significant Target Race × Participant Race interaction, F (1,28) = 0.42, MSE = 1.9, n.s. Overall, then, the pattern of PSE results suggests independent contributions of response and perceptual biases.

There are two clear patterns in the JND data shown in Fig. 4C. First, and in contrast to Experiment 1, Asian participants (M = 2.7, SE = 0.26) responded with more precision than Caucasian participants (M = 4.0, SE = 0.26), giving rise to a main effect of Participant Race, F (1,28) = 12.2, MSE = 1.1, $p < .01$, pEta = 0.3. Second, participants were more precise when the target category was the other race, giving rise to a significant Participant Race × Target Race interaction, F (1,28) = 4.8, MSE = 1.1, $p < .05$, pEta = 0.15. The main effect of Target Race was not significant, F(1,28) = 0.3, MSE = 1.1, n.s.

The median RT data are shown in Fig. 4D. The only significant effect was that of Participant Race, F(1,28) = 11.4, MSE = 664865, $p < .01$, pEta = 0.29, where Asian participants (M = 2888 ms, SE = 203 ms) took consistently longer to respond than Caucasian participants (M = 1913 ms, SE = 203 ms). Given the findings of Experiment 1, where RT varied by testing site, some caution needs to be applied in interpreting this pattern as race and location were confounded in Experiment 2. Although there appears to be a clear speed-accuracy trade-off in the data for Asian Participants – such that the more precise group with the other-race target took much longer – there was no main effect of Target Race, F(1,28) = 1.5, MSE = 664865, n.s., and no Participant Race × Target Race interaction, F(1,28) = 2.6, MSE = 664865, n.s.

## 9. Discussion

The main goal of Experiment 2 was to replicate the finding of an other-race bias when making decisions about groups of faces. Using different sets of stimuli that had been cropped and normalised to restrict information to the central features of the face, we found further evidence that other-race faces and own-race faces made different contributions to average estimates. By changing the nature of the task, we were also able to demonstrate that both perceptual and response biases affected the pattern of results, but these two factors appeared to be independent. That is, even when accounting for differences in how to respond under uncertainty, Asian and Caucasian participants appear to perceive the racial composition of our displays differently. Specifically, when judging which of the two races are more prevalent, both Participant Race groups – not just the Asian participants, as in Experiment 1 – appear to give more perceptual weight to other-race faces.

In addition to these clear shifts of the PSE, Experiment 2 also showed modulation of JNDs as a function of Participant Race. The finding of better precision when the target category is the other-race would seem to be consistent with previous visual search studies that have reported more efficient detection of other-race faces, the ORSA (Levin, 2000; Sun et al., 2013). Examination of Fig. 4C suggests that here this tuning of JND occurs for both Asian and Caucasian participants, contrasting with the search results of Sun et al. (2013) discussed in Experiment 1.

Overall, and in contrast to Experiment 1, Asian participants were more precise than Caucasian participants. While there is a hint that this might have been caused by some sort of speed-accuracy trade-off, another possibility is that the shift to only having internal facial features may have favoured the Asian participants. For example, as large variations in skin tone, eye colour and hair colour might be less useful cues in Asian cultures, their absence in Experiment 2 may have had less of an impact than on the Caucasian participants. Patterns of eye-movement data show that Asians tend to scan the centre of a face more than Caucasians, who instead fixate around different parts of the face (e.g., Blais, Jack, Scheepers, Fiset, & Caldara, 2008; Miellet, Vizioli, He, Zhou, & Caldara, 2013); Asians may therefore have been less impaired by the removal of external facial features.

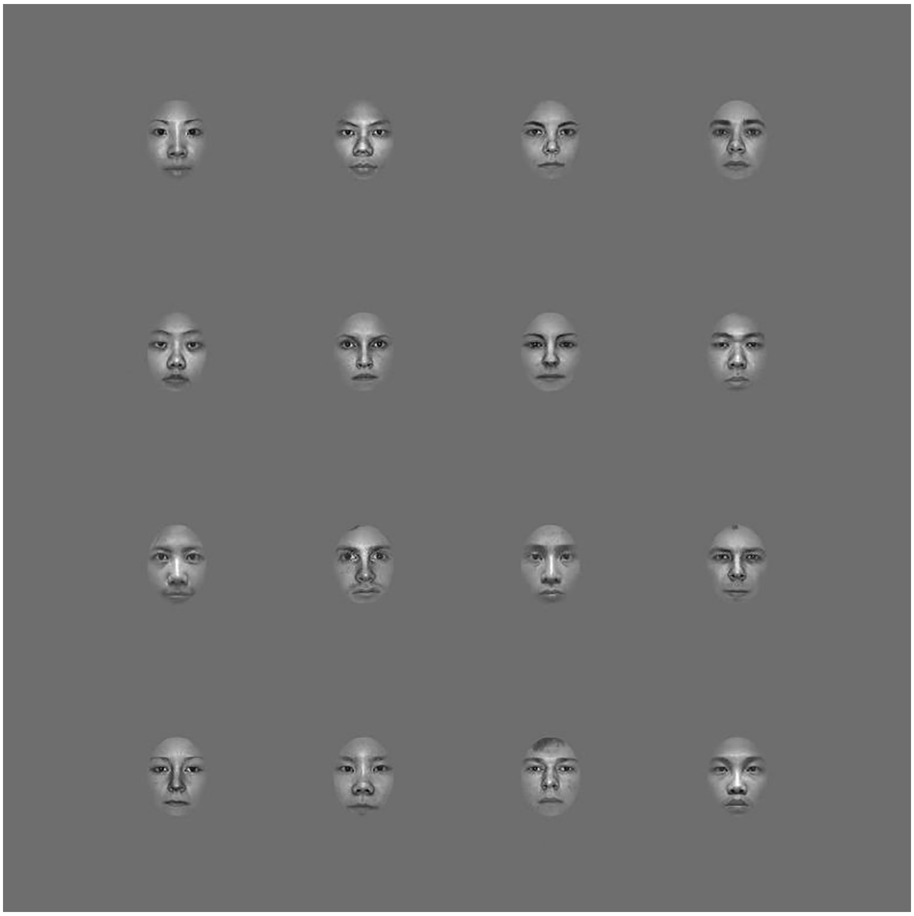

**Fig. 3.** Example stimulus array from Experiment 2 with cropped and normalised faces.

## 10. General discussion

In two experiments, we examined the ability of Asian and Caucasian participants to judge the racial composition of arrays of faces. The proportion of Asian and Caucasian faces within an array of 16 images was parametrically varied and participants had up to 4 s to make a decision on which race was in the majority. The facial arrays were animated in such a way that explicit enumeration was not possible. Participants were easily able to perform this task, suggesting that estimates of the racial composition of a group of faces could be extracted quite efficiently. Furthermore, in both experiments we found that these estimates were modulated by participant race, such that other-race faces were given more weight than own-race faces.

As noted in the Introduction, our interest in estimates about groups was inspired by previous work on ensemble representations of faces (Alvarez, 2011; de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007, 2009; Jung et al., 2017). By presenting multiple faces and preventing explicit enumeration, we created a situation where such representations could be useful in performing the task. That is, if participants could establish a representation of the "average" race present in an array – as suggested by the study of Jung et al. (2017) – they might be able to use this estimate as a basis for deciding which race was in the majority.

We have no direct evidence that this is how participants solve the current task. We chose to use different methods of presentation and a simple "majority" task, rather than trying to establish whether an average was being computed. We must therefore be quite cautious in asserting that ensemble representations of race are behind our pattern of results. A less controversial contribution that these experiments make to the literature on ensemble representation is to further highlight the potential of studying individual or group differences (Haberman et al., 2015).

If ensemble representations aren't being used here, how do participants solve the task? We introduced two dynamic manipulations to reduce the possibility that participants sub-sampled the arrays. While these manipulations meant that it would be very difficult to simply count through the displays, we cannot be certain that other types of sub-sampling were not possible. The dynamic contraction of the display was introduced to try and encourage tracking of the entire set of faces. During task development, when the contraction was not present, there was a strong subjective impression of 16 independent streams of faces, rather than a single set that was shuffling across positions. Our suspicion was that the "common fate" imposed by the contracting motion reduced the tendency to attend to only a subset of the display, which is why the contraction manipulation was used for the final experiments.

The shuffling of faces around the display provided a way to maintain the same set of faces for an extended period of time, while making it impossible to count through the individual images in the display. However, such shuffling could clearly encourage participants to focus their attention on one or two array positions and sample these positions over time (Haberman et al., 2009) rather than over space.

In our current display, attention to a single channel could provide up to 16 samples (i.e. 4000 ms total duration/250 ms per shuffle). As shuffling was random, this might not provide reliable variation that could support the levels of performance we see. However, simply attending to two or three array locations might. This would still mean that participants have to make decisions based on multiple faces, but the 4 Hz presentation rate might be slow enough for them gain a direct sense of which race is appearing more often, and use this as a basis for their decision, without the need to postulate the construction of any

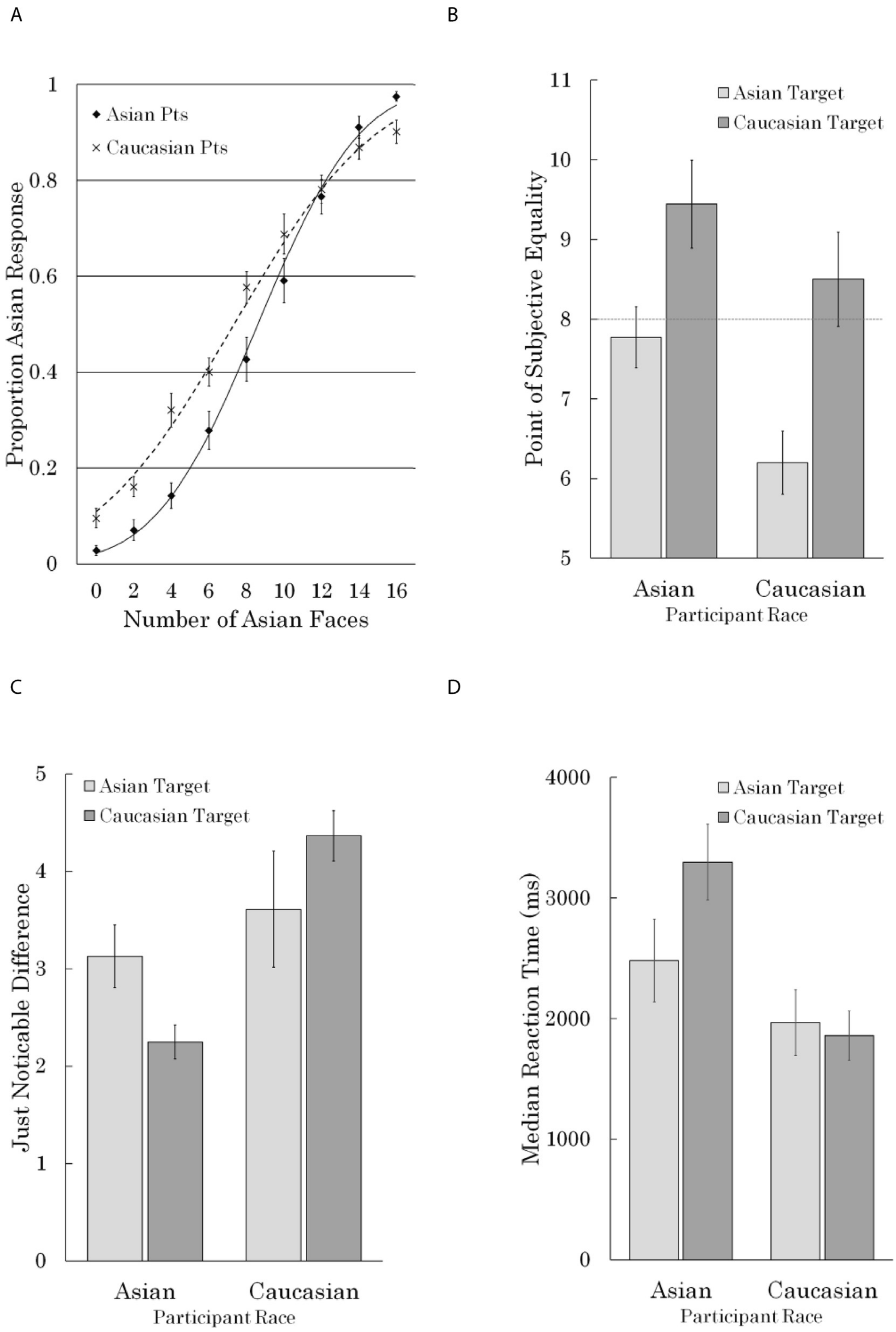

**Fig. 4.** Results of Experiment 2. A) Proportion Asian Responses collapsed across Target Race; B) PSE as a function of Target Race and Participant Race. The dotted line shows veridical equality; C) JND as a function of Target Race and Participant Race; D) Median RT as a function of Target Race and Participant Race. Error bars are standard error of the mean.

form of ensemble representation.

Few of our participants used the full available display duration, which might argue against such a strategy, but we certainly cannot rule this – or other ways of non-ensemble sub-sampling – out of the equation. In the study of Jung et al. (2017), there was evidence that differential weight was given to faces at the centre of static arrays, compared to peripheral positions, during the computation of average race. However, given the brief nature of their displays, it seems unlikely that such sub-sampling would be due to cognitive strategies involving overt shifts of attention (Wolfe, Kosovicheva, Yamanashi Leib, Wood, &

Whitney, 2015). With our displays, the central items may also have covertly attracted attention, and given our longer display durations, such biases could have resulted in the more explicit strategies mentioned above.

Although we cannot be certain of the mechanisms involved, this does not detract from our two main findings, that people can make judgements of the majority race when exposed to multiple faces, and more crucially, that those estimates are affected by their own race. As already mentioned, the first conclusion, that judgments of race frequency can be quite accurate, is consistent with previous work showing sensitivity to a range of ensemble face information, including identity, sex, emotion and race (e.g., Haberman & Whitney, 2007, 2009; Haberman et al., 2009; Jung et al., 2017). Interestingly, the recent study by Jung et al. (2017) suggests that the ability to estimate average race might be quite poor compared to other estimates of other facial dimensions. Examination of the overall pattern of results in that study (Jung et al., 2017; Fig. 3), suggests that performance is only slightly better than would be expected by a totally random process. Direct performance comparisons with other face dimensions is complicated as there have been a wide range of methods used to assess ensemble representations. However, a study by Haberman and Whitney (2010), which appears to be have been the basis of the Jung et al. (2017) task, found that individual participant standard deviations in assessing average expression ranged from between 12 and 20 emotional morph units, whereas the group standard deviations for average race in the Jung et al. (2017) study were approximately 28 morph units.

As already noted, it is difficult to directly compare our findings to ensemble representation studies, as our task was very different. We can note, however, that performance on making the race-majority decisions in our experiments does appear to be quite good. Specifically, performance was close to ceiling at the extremes of our continuum, PSEs varied quite closely around the veridical centre (i.e. 8 faces of each race), and the JNDs were relatively small, given that the minimal step size was two additional faces. Our use of photographic images that had unique identities – rather than more typical morph sequences – and/or the simple nature of the majority decision might go some way to explain why the race-estimating performance in our experiments seems quite good, while in the Jung et al. (2017), it seems quite poor. We should note, that when we reduced the cues to identity in Experiment 2, the precision of race judgements did get considerably worse, consistent with the idea that estimating race may be easier when natural cues to identity are also present.

Our second main finding – that the observer's own race influenced judgements of the group majority – has interesting implications, both for studies of ensemble representations, and for a more general understanding of the phenomenon of own-race expertise in the brain. Haberman et al. (2015) used an individual differences approach to demonstrate that the ability to construct high-level (e.g., facial expression or identity) versus low-level (e.g., orientation, colour) ensemble representations seems to be largely independent. They suggest this argues against a single "domain-general" ensemble mechanism, but rather for the existence of multiple specialised mechanisms. Our results at least imply that individual tuning of such higher-level mechanisms could depend on real-world experience/expertise. Future studies of ensemble representations might explore other group differences, such as gender, age or domain specific visual expertise (e.g., car versus bird experts).

In terms of the face recognition literature, our finding not only adds to studies of group or crowd processing, but also provides another specific example where other-race faces seem to be prioritised over own-race faces (see also, Levin, 2000; Sun et al., 2013). More generally, there has been an ongoing debate as to whether race-related effects are due to impoverished/enhanced perceptual representations (e.g., Rossion & Michel, 2011; Tanaka, Kiefer, & Bukach, 2004) versus some form of selective bias to encode other-race faces at the category-level (e.g., by race) rather than at the individual level (e.g., by identity; see for example, Bernstein, Young, & Hugenberg, 2007; Hugenberg, Miller, & Claypool, 2007). While the current study was clearly not designed to distinguish between these approaches, in the remainder of this discussion, we briefly note how each might provide an explanation for the current findings, with a view to stimulating additional research directions.

At the core of social-categorization explanations for race effects is the idea that classifying exemplars into a group "occurs quickly, effortlessly, and spontaneously upon encountering faces in most contexts" (Hugenberg, Young, Bernstein, & Sacco, 2010). Once classified, a face may then elicit a range of social cognitions associated in memory with a particular in-group or out-group (Bernstein et al., 2007). While these ideas have generally been framed within the context of individual faces, they should, in principle, apply to situations where multiple faces are presented. Indeed, one of the earliest variants of this approach – Levin's feature-selection model (Levin, 1996, 2000) – suggested that search through multiple faces could be guided (Wolfe, 1994) differentially for own and other-race faces as a function of such categorization.

Given the nature of the current task it may be the process of categorization itself, rather than any cognitive, motivational, or behavioral sequelae of categorization that is most relevant. As our task instructions explicitly provide categorization rules (i.e. Asian and Caucasian) when presented with an array of multiple faces, grouping may proceed automatically along these lines, and conceivably, be more effective for other-race, than own-race faces. We can only speculate, but if participants are more successful in grouping other-race faces into a homogenous category, this could provide a basis for the shifts in PSE that we observed in both experiments. That is, the more easily or successfully grouped category could be perceived to be more numerous. In Experiment 2, when instructions focused attention on one category versus the other, this could further enhance grouping for other-race faces, leading to the observed improvement in sensitivity (JND). Clearly, it will be important to establish whether participants have access to or can be influenced by this initial stage of categorization in these ways.

From a perceptual expertise perspective, one potential mechanism that might explain how participant race modulates performance in the current task is through the face-space model of face representation (e.g., Valentine, 1991). In this approach, faces are conceived as constituting values on multiple shape dimensions, with facial identity coded by a physical location within the space. Face-space is optimised for faces that one commonly experiences, so faces that deviate from the shape regions that best capture most of the faces around us will tend to be less-precisely represented, and therefore viewed as being more similar to each other. This provides an account for the finding that other-race faces are more difficult to discriminate from each other. In the current context, however, the higher density of other-race faces may lead to the perception that they occur with higher frequency within the set. Again, in future studies, it would be important to establish whether density and apparent frequency are related in this manner.

## 11. Conclusions

In the current work, we have demonstrated that participants are able to accurately estimate the racial composition of a group of faces under conditions that block explicit enumeration. Although further examination of this estimation ability is clearly needed, it is at least consistent with the idea that the visual system is able to compute summary or ensemble representations of race. Additionally, we have shown that such group estimates are consistently influenced by the race of the participant. This demonstrates that race-related effects operate not only at the level of individual face processing but also when making decision about groups of faces. Our initial findings could potentially be explained from within both perceptual-expertise and social-categorization accounts, but future studies using similar methodology may provide greater clarity on this issue.

## References

Allport, G. W. (1954). *The nature of prejudice.* Oxford, England: Addison-Wesley.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162.

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science, 18*(8), 706–712.

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PloS one, 3*, e3022.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Chen, L. F., & Yen, Y. S. (2007). Taiwanese facial expression image database. [http://bml.ym.edu.tw/download/html]. Brain Mapping Laboratory. Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology, 62*(9), 1716–1722.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432–446. http://dx.doi.org/10.1037/xge0000053.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(11), 1–13 http://journalofvision.org/9/11/1/, doi: 10.1167/9.11.1.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 718–734.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72*(7), 1825–1838.

Hugenberg, K., Miller, J., & Claypool, H. M. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology, 43*(2), 334–340.

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review, 117*(4), 1168.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*, 313–323. http://dx.doi.org/10.1016/j.cognition.2011.08.001.

Jung, W., Bülthoff, I., & Armann, R. G. M. (2017). The contribution of foveal and peripheral visual information to ensemble representation of face race. *Journal of Vision, 17*(13), 1–12. http://dx.doi.org/10.1167/17.13.11 11.

Kleiner, M., Brainard, D., & Pelli, D., 2007. What's new in Psychtoolbox-3? Perception 36 ECVP Abstract Supplement.

Laurence, S., Zhou, X., & Mondloch, C. J. (2016). The flip side of the other-race coin: They all look different to me. *British Journal of Psychology, 107*(2), 374–388.

Levin, D. T. (1996). Classifying faces by race: The structure of face category. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1364–1382.

Levin, D. T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General, 129*(4), 559–574.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology, 13*(4), 330–334.

Miellet, S., Vizioli, L., He, L., Zhou, X., & Caldara, R. (2013). Mapping face recognition information use across cultures. *Frontiers in Psychology, 4*, 34.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36*, 630–633.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Prins, N., & Kingdom, F. A. A., (2009). Palamedes: Matlab routines for analyzing psychophysical data. http://www.palamedestoolbox.org

Rossion, B., & Michel, C. (2011). An experience-based holistic account of the other-race face effect. *The Oxford handbook of face perception*, 215–243.

Sun, G., Song, L., Bentin, S., Yang, Y., & Zhao, L. (2013). Visual search for faces by race: A cross-race study. *Vision Research, 89*, 39–46.

Tanaka, J. W. (2013). Recognition of own- and other-race, gender and species faces [Special issue, Introduction]. *Visual Cognition, 21*(9–10), 1077–1080.

Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. *Cognition, 93*(1), B1–B9.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43A*, 161–204. http://dx.doi.org/10.1080/14640749108400966.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*(1), 105–129. http://dx.doi.org/10.1146/annurev-psych-010416-044232.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Attention, Perception, & Psychophysics, 63*(8), 1314–1329.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*(3), 671–684.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review, 1*(2), 202–238.

Wolfe, B. A., Kosovicheva, A. A., Yamanashi Leib, A., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision, 15*(4), 1–13. http://dx.doi.org/10.1167/15.4.11.