Do individual differences in face recognition ability moderate the other ethnicity effect?

Michael Jeanne Childs[1], Alex Jones[1], Peter Thwaites[2], Sunčica Zdravkovic[3], Craig Thorley[4], Atsunobu Suzuki[5], Rachel Shen[6], Qi Ding[6], Edwin Burns[7], Hong Xu[8], and Jeremy J. Tree[1]

[1]Swansea University, United Kingdom

[2]Keimyung University, Korea

[3]University of Novi Sad, Serbia

[4]James Cook University, Australia

[5]University of Tokyo, Japan

[6]Nanjing University, China

[7]Edge Hill University, UK

[8]Nanyang Technological University, Singapore

Word count: 8775

Author's note:

Analysis and results using the Caucasian sample data from this study was presented in EPS Bournemouth Conference (July 2019).

The authors have no conflicts of interests to disclose. Please find the coding and raw data on osf.io/bwhtg.

Correspondence concerning this article should be addressed to Professor Jeremy Tree, Department of Psychology, Swansea University, United Kingdom, SA2 8PP

E-mail: j.j.tree@swansea.ac.uk

Abstract

Individuals are better at recognizing faces from their own ethnic group as compared to other ethnicity faces – the *other-ethnicity effect* (OEE). This finding is said to reflect differences in experience and familiarity to faces from other ethnicities relative to faces corresponding with the viewers' ethnicity. However, *own-ethnicity* face recognition performance ranges considerably within a population, from very poor to extremely good. In addition, within-population recognition performance on other-ethnicity faces can also vary considerably with some individuals being classed as '*other ethnicity face blind'* (Wan et al., 2017). Despite evidence for considerable variation in performance within population for faces of both types, it is currently unclear whether the magnitude of the OEE changes as a function of this variability. By recruiting large-scale multinational samples, we investigated the size of the OEE across the full range of own and other ethnicity face performance whilst considering measures of social contact. We find that the magnitude of the OEE is remarkably consistent across all levels of within-population own- and other-ethnicity face recognition ability, and this pattern was unaffected by social contact measures. These findings suggest that the OEE is a persistent feature of face recognition performance, with consequences for models built around very poor, and very good face recognisers.

Keywords: other-ethnicity effect, face memory, individual differences, face recognition, developmental prosopagnosia, super-recognisers

Public Significance Statement: This study provides an important new piece to the puzzle of understanding a fundamental characteristic of human face processing that is the other-ethnicity effect. We found that this phenomenon is universal and 'fixed' across the spectrum of individual face processing ability across nations.

Do individual differences in face recognition ability moderate the other ethnicity effect?

The face plays a central role in human social interaction. Typically, from a young age, we are able to identify familiar faces which aids in survival and attachment (Barrera & Maurer, 1981), and as we age, our ability to recognise faces in different contexts allows us to distinguish between familiar and unfamiliar faces which has an impact on our interpersonal relationships (Gobbini et al., 2004). A consistently reported phenomenon in facial recognition is that typically developing samples are generally better at recognizing faces from their own ethnicity compared to other ethnicities; also known as the other-ethnicity effect (OEE; Malpass & Kravitz, 1969; McKone et al., 2012). A well-known theoretical account of this effect is posited by *perceptual expertise theory*, which suggests that the OEE reflects a lack of experience in seeing and encoding other-ethnicity faces. Supporting evidence comes from infant studies, where 6-9 month-old infants were shown to be able to discriminate between own-ethnicity and other-ethnicity faces (Sangrigoli & De Schonen, 2004; Anzures et al, 2013, Kelly et al., 2007). Training studies, where participants show reduced OEE after training with other-ethnicity faces (Lebrecht et al., 2009) also support this notion.

The OEE and contact

A key factor that is claimed to impact one's performance with faces of different ethnicities relates to the amount of social contact they have with certain groups. The *contact hypothesis* posits that the higher the contact an individual has with faces of a particular ethnicity, the more accurate they are at recognizing members of that group (Goldstein & Chance, 1985). For example, Zhou et al. (2019) demonstrated that Caucasians and East Asians born and raised in the wider Toronto area had comparable face recognition abilities for Caucasians and East Asian faces (i.e. East Asians born in the Toronto area did not display an OEE for Caucasian faces). In addition, length of exposure to Caucasian faces moderated

the OEE for East Asians (i.e., the longer they had lived in Toronto, the smaller the OEE). In general (although see Ng & Lindsay, 1994, Harvey, 2010, and MacLin et al., 2004), studies investigating the role of contact (both in geographical and self-report) in face recognition show that as contact increases so the magnitude of the OEE can diminish (see Table 1 for summarised findings). However, it is particularly noteworthy that although contact can diminish the magnitude of the OEE, it often does not eliminate this effect completely (De Heering et al., 2010; although see Estudillo et al., 2020).

[INSERT TABLE 1 HERE]

Variations in individual face processing performance for own- and other-ethnicity faces

In Table 1, we have provided a summary of several studies of the OEE that explored the degree to which the effect is impacted by social contact (e.g., high 'contact' group versus low 'contact' group), and a number have shown that across groups, the magnitude of the OEE can indeed vary. But this often masks the fact that *within* groups there is often considerable variance in individual ability with own-ethnicity faces – where it is often implicitly assumed that *own-ethnicity* face performance (i.e., baseline face recognition ability) across two samples of the same population (e.g., two UK Caucasian populations) is quite homogeneous, such that between group differences are driven by other variables (such as social contact). There is now a great deal of evidence that suggests that the range of own-ethnicity face recognition accuracy across individuals for a particular population can be substantial (i.e., several standard deviations), and thus raises an important question – might the magnitude of the OEE change as a function of this variability? One approach to exploring this question is to focus on the performance of sub-populations linked to the 'extremes' of this distribution of own-ethnicity (baseline) face recognition ability – namely, on those who are performing very

poorly (developmental prosopagnosia) or those performing extremely well (super recognisers). The logic being, if individual variability in baseline face recognition does impact on the emerging OEE, one might expect differences between sub-populations – and we will discuss this work now. Our study takes a novel approach, however, by exploring the degree to which the magnitude of the OEE varies across *the full* distribution of base level face recognition ability, and thus considers this issue in the widest possible sense (more below).

People with *developmental prosopagnosia* (DP) have impairments in recognizing own-ethnicity faces despite having normal intelligence and an absence of brain injury (Bate et al., 2019; Burns, Bennets, et al., 2017; Burns, Martin, et al., 2017; Burns et al., 2014; Jackson et al., 2017). Interestingly, in our experience, DPs often report anecdotally *"all faces look the same to me…."* and "*I often confuse two different people who I know that look similar…"* (Bate & Tree, 2017).[1] This raises an interesting question – perhaps poor base-level face recognition ability emerges because of a general inability to draw from one's visual experience when learning faces? - that is, *despite* high familiarity/experience with own ethnicity faces, performance remains poor. If this is true, then we might expect poor face recognisers to do equivalently (with no OEE) across all ethnicities of faces ("*all faces look the same…")*, since high visual experience gives them little benefit at all. However, perceptual studies (DeGutis et al., 2011; Cenac et al., 2019) have found that DPs as a group demonstrated an OEE. A recent study by Cenac et al. (2019) looked at facial recognition abilities of Caucasian controls and DP participants using a sequential matching task (with Caucasian, East Asian, and Black ethnicities). All participants were matched on measures of social contact with other-ethnicity faces (i.e., minimal contact with people from East Asia and Black backgrounds). Cenac et al. (2019) concluded that DPs in their sample did not have

[1] In addition, very recently a DP volunteer in our lab mentioned that he had confused his girlfriend with his best friend's girlfriend because they had superficial physical similarities (similar height, build, hair colour/style and clothing), despite the fact that one was Asian and the other Caucasian.

disproportionately poorer performance for other-ethnicity faces relative to controls. However, their findings could not speak to the issue of whether the OEE was present across both groups because this study did not find an overall OEE for either group (*despite* the low degree of contact). Indeed, the data reported by Cenac et al. (2019) illustrated a trend towards an inverted OEE – with controls and DPs better at matching other-ethnicity faces. It remains unclear why this occurred, but might reflect the deliberate increased variability of the other-ethnicity faces in their stimuli (all computer generated), which may have made the other-ethnicity faces easier to discriminate. In any case, no typical OEE was reported using their paradigm, which may be problematic with respect to interpreting their findings. Putting this issue aside, their findings suggest that DP cases are largely worse than controls for *both* own and other ethnicity faces on testing of face perceptual matching.

Conversely, people dubbed *super recognisers* (SR), are reported to do extremely well with own-ethnicity faces (Ramon et al., 2019). These individuals may thus show a general 'boost' to recognition performance for faces of a variety of ethnicities (outside their own), such that for them the OEE may be relatively diminished. Alternatively, SRs may still show an own-ethnicity face advantage despite their generally excellent face recognition abilities. Similar findings from Bate et al. (2018) and Robertson et al. (2019) independently provide evidence for the latter pattern using various face memory and face matching tasks, which show that while SRs outperformed a matched sample on respective tests (i.e., better performance with *both* own- and other-ethnicity faces), a similar OEE size was found across the two groups. This suggests that even when base-level face recognition performance is extremely good, an advantage remains for own-ethnicity faces.

Thus, there is preliminary evidence that the OEE persists at the 'extremes' of *own-ethnicity* recognition performance within a given population – when this is considered via a comparison of performance between population sub-groups. However, there remains an

additional pattern of 'extreme' within-population individual performance to be considered; namely, extremely poor *other-ethnicity* performance. Given the fact that within a population there is a distribution of performance with own-ethnicity faces, an assumption is that a similar distribution exists for individuals with other-ethnicity faces, and that these distributions are correlated, moving together. However, it may also be possible that there are individuals who have very poor performance with other-ethnicity faces despite good own-ethnicity face performance, akin to an exaggerated form of OEE – a pattern dubbed '*other ethnicity blindness'*.

To explore this issue, Wan et al. (2017) tested samples of both Caucasian and Asian participants on the Australian and Asian-CFMTs – in order to identify such 'extreme' poor performers they used absolute cut-off scores for each test (i.e. mean accuracy minus 2 standard deviations; SD). Participants who scored lower than 2SDs below the mean on their own-ethnicity face memory test were excluded to rule out the influence of general poor facial recognition ability (i.e., developmental prosopagnosia). Caucasian participants who met the criteria for 'other ethnicity blindness' were thus identified using a cut-off from the Asian participants' sample on Asian-CFMT (and vice versa for Asian participants) – and under this criteria, it was found that 8% (N=36) of the sample performed lower than 2SDs below the mean. It was further argued that this selectively extremely poor facial recognition for other-ethnicity faces was neither due to lack of effort, nor poor general facial recognition ability, and that the level of contact may influence such cases. However, we would point out that this study only used one CFMT test to 'diagnose' participants who were other-ethnicity face blind. Typically, two or more tests are used to 'diagnose' DP (i.e. own-ethnicity face blindness); thus it is unclear whether the cases identified would continue to meet criteria for 'other-ethnicity face blindness' if other tests had been used - given the possibility of regression to the mean (discussed below). Nonetheless, this work suggests if we consider

within population individual variance on *other-ethnicity* face recognition, it may be the case that the magnitude of the OEE varies across the distribution, where it may be being magnified at one extreme.

However, in all these studies of the OEE with individuals who are in the 'extremes' of own- or other-ethnicity face recognition ability, the approach has been to compare a (often quite small) sample of their 'extreme' group with another sample that comprised the rest of the population. A key criticism of this practice is that it involves the use of an arbitrary cut-off criteria score (2 SDs below average on a key test, as described above) for group categorisation, which likely does not reflect qualitative differences in performance. In other words, participants with performance either side of such a cut-off (i.e., 2.02 SD below average versus 1.98 SD below average) may artificially imply key group differences even when the performance between individuals may not be significantly different. This is a key motivation for the current study's novel approach – since it ensures explicitly that we did not group the participants into categories, but rather considered performance across the *entire* distribution (i.e., at all levels of performance from extremely poor to extremely good) – and thus we can ask (for the first time) whether the magnitude of the OEE remains equivalent across *all* levels of performance in a given population. It is important to note that although there are studies which looked at all levels of recognition performance in a population using both own- and other-ethnicity face recognition tasks, (e.g. Robertson et al., 2019 and Horry et al., 2015), they do not explicitly measure the magnitude of OEE across the whole of the population – in Robertson et al.'s case, they only made comparisons of OEE magnitude for super-recognisers and controls, and in Horry et al.'s case, they only reported the correlation of own- and other-ethnicity face recognition performance – and therefore do not necessarily touch upon this matter.

Furthermore, not only do we consider the question of the size of the OEE across *own-ethnicity* face performance, we also explore the same issue from the position of *other-ethnicity* face performance. To our knowledge, this is the first study to use an out-group face ability measure as a predictor of OEE, which opens another avenue for us to understand this effect further.

Finally, given we are interested in the *universality* of the magnitude of the OEE across individuals in a population, we also sought to explore this issue across a number of different nations with populations that were either largely Caucasian (UK, Australia, and Serbia) or largely Asian (China, Japan, S. Korea, and Singapore), and thus the multi-national nature of our sample would allow us to investigate OEE in a more extensive manner.

Exploring within population individual variation in face recognition

It is noteworthy that a potential criticism of some of the previously discussed research on sub-groups of 'extremes' of individual performance is that they largely studied face *perception* (i.e., face matching), rather than face recognition. This is despite the fact that group-based studies of the OEE (see Table 1) have often focused on face *recognition*. To address this issue, another key motivation for the current study was that it sought to focus on individual variation within a population on measures of face recognition performance. In order for us to achieve this objective it was important for us to use a well-validated measure of face recognition ability – and so we selected the Cambridge Face Memory Test (CFMT). In this case, we used three well-established versions: Boston (Duchaine & Nakayama, 2006), Australian (McKone et al., 2011), and Asian (McKone et al., 2012). In Table 2, we summarise a number of studies that used versions of the CFMT to investigate the OEE – importantly, in all cases the studies report a robust OEE (Cohen's *d* effect sizes between 0.5 – 1.24). In addition, because of its established validity and reliability, the CFMT has been used in a great range of individual differences work relating to face recognition over the last

fifteen years (see Wilmer, 2017 for a comprehensive review). Thus, we have confidence that the CFMT is a robust tool for our current purposes.

[INSERT TABLE 2 HERE]

An impetus for using the CFMT and the final motivation for the current study relates to the fact that it has three different versions (mentioned above) – and thus we would be able to utilise a CFMT test (e.g., the CFMT Boston) as an *independent* measure of individual face recognition memory performance from those used to traditionally capture and calculate the OEE (e.g., CFMT Australian versus CFMT Asian). This enables us to consider an important potential confound - *regression to the mean* (that is, individual performance can vary around its "true mean", such that an extreme high or low score may naturally move on its second measurement). Put simply, if a key group of interest (DPs, super- recognisers, or cases of other-ethnicity blindness) is initially selected via 'extremely' poor scores on one measure (own-ethnicity face recognition), it is likely these same participants might be less poor on a second measure of face recognition because of regression to the mean. Therefore, the observed differences between two tests could be simply due to this phenomenon when the same test is used as the classifier *and* a comparator. Having a third face recognition memory measure that would provide an independent measure of face recognition memory from that used to compute the OEE was thus extremely useful, and the three well-established variants of the CFMT made it ideal for our purposes.

The fact that the CFMT has three variants also made it ideal for the current study given we sought to recruit large samples of both Caucasian and Asian participants. The current study aims to use these three CFMT variants in testing these different populations in

order for our analyses to ask two different, but related questions. Firstly, for the Caucasian sample, our independent measure of face recognition memory is a Caucasian stimulus set (the 'Boston' CFMT), and so we will be determining whether the magnitude of the OEE varies as a function of individual ability for *own-ethnicity* faces. For the Asian sample, our independent measure of face recognition memory is the same Caucasian stimulus set ('Boston' CFMT), and so in this case we will be determining whether the magnitude of the OEE varies as a function of individual ability for *other-ethnicity* faces. Thus this work will consider the OEE in a manner never yet attempted – it will ask does the size of the OEE vary across a given population when considered either across the distribution of own-ethnicity performance (in three different large Caucasian samples) or across the distribution of other-ethnicity face recognition performance (in four different large Asian samples).

Study Aims and Implications

In summary, the primary aim of this study was to investigate the OEE across within population distributions of own- and other-ethnicity face recognition performance. For the most part, previous work has often focused on 'extremes' of performance with either own-ethnicity (i.e., very poor performers – DP or very good performers – super recognisers) or other-ethnicity faces (i.e., other-ethnicity blindness), and we have raised various methodological issues with several previous studies. Instead of (somewhat arbitrary) comparisons of performance across sub-groups of a given population, we have taken the approach of considering the pattern and magnitude of the OEE across *all levels* of face recognition ability. Thus allowing us for the first time to determine whether this OEE pattern might in some way vary in size as a function of within population individual variance for *own-ethnicity* faces on the one hand and for *other-ethnicity* faces on the other hand (whilst also controlling for social contact).

Our findings will have interesting implications – if it is determined that the magnitude of the OEE for individuals in a given population is in fact impacted by their relative performance as indexed at baseline by an own- or other-ethnicity face measure, this has consequences for future studies of the OEE going forward (since they must take this into account). However, if it is determined that the magnitude of the OEE remains constant across both distributions of performance, this would provide interesting evidence of the universality of the OEE in face recognition performance. Thus we believe that understanding the degree to which the OEE is impacted by within population individual variation will speak both to previous work on the OEE that has been undertaken (see Tables 1 & 2) and to studies of group comparisons of the OEE that have focused on comparisons with participants who perform at the 'extremes' of these distributions.

Method

Participants

Eight hundred and fifty-two participants (largely undergraduate students - see Table 3) were recruited from universities in their respective countries. Participants were recruited in their respective universities as part of their Psychology course requirement. 28 participants did not complete the study and were therefore their data were excluded from analysis (N=824). Informed consent was acquired prior to the start of the experiment. All participants had normal or corrected to normal vision during test completion. As we sought to consider the OEE and influence of contact, recruiting solely from one country could mean that we are not able to capture differences in the level of contact. We therefore sought to recruit across nations for which we may assume there are varying levels of contact with other ethnicities (e.g. UK has more diverse population than Serbia, and a rural University in China would have less diverse population than South Korea and Japan). Additionally, recruiting from different

countries of similar ethnic groups would give us a more diverse sample and increase the generalisability of the findings.

[INSERT TABLE 3 HERE]

Statement of Ethics

All participants gave written consent forms and were compensated with study credits for participating. This study was approved by the Swansea University Ethics Committee and followed the Declaration of Helsinki (World Medical Association, 2009).

Materials

Cambridge Face Memory Test (CFMT) Versions

To estimate the OEE, we employed face recognition tasks that utilise faces from different ethnicities. In this case, we used three well-established versions. First was the original 'Boston' task, which primarily has faces from Harvard University with South European or Middle Eastern features (Duchaine & Nakayama, 2006). Internal reliability (IR) for this version was reported to be between .86-.90 for Caucasian participants (Bowles et al., 2009; Wilmer et al., 2010; McKone et al., 2012; DeGutis et al., 2013) and .94 for Asian participants (McKone et al., 2012). Second was the 'Australian', which has a combination of primarily Caucasian British-ethnicity faces from Australia, New Zealand, and Scotland (McKone et al., 2011). IR for this version was reported to be between .88-.89 for Caucasians (McKone et al., 2011; Horry et al., 2015) and .85 for Asians (Horry et al., 2015). Finally, the 'Asian', which primarily has Han-Chinese faces (McKone et al., 2012). IR for this version was reported to be .88-.90 for Asian participants (McKone et al., 2017; Horry et al., 2015) and between .77-.89 for Caucasian participants (Horry et al., 2015; McKone et al., 2012;

DeGutis et al., 2013). Overall, these studies demonstrate that the different versions of the CFMT are reliable in detecting OEE, as given by the high internal reliability found from the tasks as well as the similarity in difficulty levels across the tests (McKone et al., 2011; McKone et al., 2012).

All CFMTs followed the original procedure outlined by Duchaine and Nakayama (2006), shown in Figure 1. All faces were greyscale images of males, with hair cut-out. All versions had three phases. (1) Learn (18 trials; three target faces) – participants were shown the target faces in three views (left, front, right) and were asked to identify the target in a triad (one target and two distractors). (2) Novel (30 trials, six target faces) – participants were shown the target faces in different lighting or viewpoint in a triad with two distractors. Finally, (3) Noise (24 trials) was similar to the Novel phase, but with Gaussian noise added to increase the difficulty of the task. Between each phase, all six target images were presented in front view to the participants for 20 seconds as a reminder. For each test version, accuracy of identifying the target faces was recorded for every phase and they were summed to obtain total accuracy (72 trials). Therefore, the higher the score, the better one's facial recognition ability. Each of the CFMTs was presented to participants in a set of three different orders (balancing which CFMT was seen first), and in line with previous findings (McKone et al., 2012), no significant differences between presentation orders was found (see Supplementary Materials).

[INSERT FIGURE 1 HERE]

Social Contact Scale (Walker & Hewstone, 2006)

To measure self-reported contact, we used a ten item, 5-point Likert questionnaire. Item 1 asked how many people from the other ethnicity participants knew - Up to 2, Up to 5,

Up to 8, Up to 10, Up to 12. Items 2-5 pertain to the social component of the questionnaire, which asked how much contact participants have with the other ethnicity, e.g. 'I often spend time with East Asian (White) people, using the following scale: *strongly agree, sort of agree, not sure, sort of disagree, strongly disagree*. Items 6-10 pertain to the individuation component, which asked participants how often they engaged with the other ethnicity, e.g. 'I have looked after or helped a South Asian (White) friend when someone was causing them trouble or being mean to them', using the following scale: *very often, quite often, sometimes, hardly ever and never*. The latter two subscales were scored so that lower values indicate higher levels of the measure, while the first subscale simply counts the number of people from other ethnicity group the person knows. To make analyses more straightforward, we reverse scored Social and Individuation components.

Table 4 presents the contact scores for this study; it is clear that average contact scores for both measures were largely quite low (perhaps surprising given our sampling across different countries), and variability in contact within populations was also reasonably small (social and individuation contact – see Walker & Hewstone, 2006). Therefore our contact measure was collapsed – and we used overall mean contact scores for the subsequent analyses, with higher scores representing more contact (individual components are more fully explored in the Supplementary Materials).

[INSERT TABLE 4 HERE]

Procedure

Participants were recruited in their respective Universities as part of their Psychology course and completed the study in the laboratory. Participants were provided with a Participant Information Sheet, and informed consent was acquired prior to commencing of

the study. All participants completed the Social Contact questionnaire (Walker & Hewstone, 2006) before starting the battery of CFMTs.

The computer tasks were presented using a bespoke programme constructed by the department's software technician, following the methods outlined for the CFMT (Duchaine & Nakayama, 2006). The order of CFMTs was counterbalanced for each participant to reduce order effects (see Supplementary Materials for further analysis). Following completion, participants were thanked for their time and awarded course credits.

Data Cleaning

28 participants did not complete all tasks and therefore their data were not included in the final dataset used for our analysis. In addition, all test scores for individuals were inspected and all were at above chance performance of 24 (Cho et al., 2015), therefore no further data exclusions were made.

Design and Analytic Strategy

To address our questions, we built a statistical model that allows us to simultaneously estimate the size of the OEE, the effect of social contact, and independent own- or other-ethnicity recognition performance on CFMT scores. Importantly, it allows us to estimate the interactions between these variables, revealing how the magnitude of the OEE is affected by other variables. For example, it is possible that the size of an individual's OEE depends on their own-ethnicity or other-ethnicity recognition ability, their amount of social contact, or both. Here, we build two separate models to test these effects in our Caucasian (n = 400) and Asian (n = 424) sample of participants, respectively.

To estimate these effects, we utilised a linear mixed regression model, with three main predictors and the full set of interactions between them. Our model structure is as follows, with exposition on the predictors and their interpretation:

$$Y_{si} = (\beta_0 + S_0) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3$$

Where B_0 represents the own-ethnicity test scores, S_0 pertains to participant error, X_1 represents the difference between own-ethnicity and other-ethnicity test score, i.e. OEE, X_2 represents Boston scores, and X_3 represents average social contact scores.

For both models, we z-scored standardised both the Boston CFMT and Social Contact scores across all the available data (separately for Asian and Caucasian participants). This meant that our models are easily interpretable. The intercept, β_0, represents the average score on the reference-coded CFMT task (for the Caucasian model, the Australia CFMT, and for the Asian model, the Asia CFMT). The random intercept, S_0, is estimated per-participant, and accounts for the fact that the Asian and Australia CFMT scores are sampled from the same individual. They thus represent the offset from the overall intercept. Models were estimated using lme4 in R (Bates et al., 2015).

The dependent measure here are the scores on the Australia and Asia CFMTs, collapsed into a single vector of scores, nested within participants as a repeated measure. For example, the *i*th score may represent the score on the Australia CFMT for participant *s*. We aimed to predict these scores as a function of the following inputs.

The coefficient $\beta_1 X_1$ is the effect of a categorical variable that coded the CFMT task that a given score was taken from – that is, the Australia or the Asia CFMT. For our model fitted to Caucasian data, the Australia CFMT was coded with zero (i.e., was designated the reference category) and the Asia CFMT coded as one. For the model fitted to Asian data, this was reversed. This has the effect of making the own-ethnicity CFMT task the baseline or reference measure. We labelled this the Face Memory Test coefficient (FMT). Importantly,

when estimated, this coefficient represents the OEE, measuring the differences between the scores of the Australia and Asia CFMTs. A useful conceptualisation of this coefficient, which is the crux of our model, is that it allows us to fit two slopes simultaneously to the data – one for the Australian CFMT scores, and one for the Asian CFMT scores. For example, these two slopes can run parallel to one another or move in different directions, if an interaction is present. This allows us to negate issues of difference scores or the use of residuals that are common, as they have undesirable statistical properties and bias estimates of effects (McElreath, 2020; Freckleton, 2012; DeGutis et al., 2013). It also ensures the difference between the CFMT tasks is estimated simultaneously with other predictors, and thus is not the same as simply subtracting one CFMT from the other.

The coefficient β_2X_2 represents the scores on the Boston CFMT. For our Caucasian participants, this is taken as an independent own-ethnicity performance measure that may predict the dependent measure, and conversely for our Asian participants, this coefficient represents an independent other-ethnicity performance measure. The coefficient β_3X_3 represents the average scores on the Social Contact scale, with higher values representing more contact with individuals of different ethnicities.

It follows that the coefficient of $\beta_4X_1X_2$ represents the interaction between FMT and scores on the Boston CFMT. Thus, this coefficient can represent a different slope between the Australia and Asia CFMTs. If, for example, individuals with higher own-ethnicity recognition (or other-ethnicity, for Asian participants) ability exhibit a smaller OEE, this coefficient would represent such an effect, with the slopes for the Australia and Asia initially being far apart but coming closer together as Boston scores increase. Very similarly, the coefficient of $\beta_5X_1X_3$ represents the same effect but with Social Contact scores – if individuals with higher contact exhibit a smaller OEE, this coefficient would represent this difference. The coefficient of $\beta_6X_2X_3$ allows individuals with higher scores on the Boston and

Social Contact measures to have different scores on *either* the Asia or Australia CFMTs, which is of less theoretical interest. However, this term is included in the model, as we wish to test the three-way interaction ($\beta_7 X_1 X_2 X_3$) between FMT, Boston, and Social Contact – that is, whether individuals with high or low scores on *both* the Boston CFMT and Social Contact measure exhibit a larger or smaller OEE. Interaction variables are taken as the multiplication of their components.

We conducted a power analysis via simulation to estimate the smallest effect we could detect with our design, which was between .20 and .25 for each coefficient (i.e., a one unit change in the predictor equates to a .20-.25 unit change in Asia or Australia CFMT scores) at 80% power, which is a very small effect (see Supplementary Materials for full details).

Results

Descriptive Statistics

Table 5 presents the mean scores and standard deviations for each of the CFMT versions for each country cohort and collapsed by ethnicity. Overall, Caucasian participants scored higher than Asian participants in the two Caucasian versions of the test, while Asian participants scored higher in the Asian version of the test.

(INSERT TABLE 5 HERE)

Reliability Analysis

To determine the internal reliabilities of our measurements we undertook several analyses. Firstly, our selection of the CFMT tests was (as we established earlier) largely

motivated by previous work that has established their high measurement reliability. Nonetheless, we checked the internal reliabilities for each of the CFMT versions across our sample, and determined Cronbach's alpha values: Boston CFMT *a* = .917, Australia CFMT *a* = .873; and Asia CFMT *a* = .846. Split into the two ethnicity groups, our analysis yielded similar *a* values, for Caucasians: Boston CFMT a = .933, Australia CFMT *a* = .863, and Asia CFMT *a* = .820; and for Asians: Boston CFMT *a* = .883, Australia CFMT *a* = .851, and Asia CFMT *a* =.843. These correspond with the reports of internal reliability values in other studies (see Horry et al., 2015; McKone et al., 2012), confirming that the use of these CFMT versions was appropriate.

However, although each independent test shows high internal reliability, the OEE, which is derived in our models as a covariate-adjusted difference between the two measures, may not be (Sunday et al., 2017, Ross et al., 2015). No study has yet investigated the internal reliability of the OEE itself, and thus it remains an open question as to whether this measurement may in fact be far noisier than has previously been assumed, and thus throwing doubt on findings focused on individual performance (e.g., the lack of interactions found between OEEs and other variables may be due to the noise in the measurement). However, it is also important to note that the linear mixed model approach used in our analysis can closely incorporate individual performances on the CFMTs by estimating individual offsets from the global intercept, which was both a major motivation and advantage of choosing the analytical approach we presented here.

In order to explore the internal consistency of the OEE, we first divided the items into the phases as described by Duchaine and Nakayama (2006), i.e. Learn (items 1-18), Novel, (items 19-48), and Noise (items 48-72). Within these phases, we randomly split the items into two equal size groups – e.g the first nine random items from Learn phase were labelled *Learn 1*, the first fifteen random items from Novel phase were labelled *Novel 1*, and the first twelve

random items from Noise phase were labelled *Noise 1*, and so forth. Using a bootstrap resampling approach, we created these random splits 9,999 times, and summed the scores within the each split across the different phases, which created composite scores for the half of the test, i.e. Learn 1, Novel 1, and Noise 1 were collapsed together to make a composite score - Split 1.

Using the split-halves mentioned above, we took the difference for each of the test halves between the corresponding own-ethnicity and other-ethnicity score for our samples, e.g. for Asian samples, we used Asia Split 1 – Australia Split 1 and for Caucasian samples, we used Australia Split 1 – Asia Split 1 to create an OEE 1 score, and so on. Using Spearman-Brown correction, we analysed the reliability of the OEE scores for each of the split pairs, generating a distribution of split-half reliability coefficients. We tested this within the full sample, and within each participant ethnicity subsample. The means were highly similar, 0.64 for the full sample, 0.65 for the Caucasian sample, and 0.63 for the Asian sample. The distributions are shown in Figure 2.

[INSERT FIGURE 2 HERE]

Although the mean *a* values for the OEEs are lower than that of the CFMT measures on their own, they are still within acceptable levels (Ursachi et al., 2015). Nonetheless, it is striking that the OEE measure is indeed lower in internal reliability, and this demonstrates that although our individual measures did have very high reliability, the *difference* between these measures (the reported OEE) was lower. This indicates for the first time that work exploring individual differences and the OEE, must utilise very reliable face recognition measures across ethnicity and report internal reliability scores for the OEE they have determined.

Caucasian Model

The estimated coefficients for the model fit to the data from Caucasian participants are shown in Table 6. Only two predictors were statistically significant. The first was the FMT, which estimated the difference between the Australia (coded zero) and Asia CFMTs, b = -3.97, $t(395.99) = 9.76$, $p < .001$, thus representing a significant OEE effect. This is directly interpretable as the Asia CFMT having, on average, a lower score than the Australia CFMT by 3.97 points. Second was the Boston CFMT predictor, which here represented an independent measure of own-ethnicity performance, $b = 5.03$, $t(731.19) = 14.98$, $p < .001$. Thus, as individual scores on the Boston CFMT increased by one standard deviation, on average, scores on Australia CFMT increased by 5.03 points. There was no significant effect of social contact, and notably, we observed no significant interactions between the FMT predictor or the Boston predictor. This indicates that while the scores on the Asia CFMT are lower than the Australia CFMT, the slope changes by more or less the same amount for each with increasing own-ethnicity recognition ability (measured by the interaction between the FMT and Boston coefficient, $b = 0.16$) or social contact (measured by the interaction between FMT and social contact coefficient, $b = 0.55$). The interaction between all three predictors was also not significant. Despite this, the variance explained by the fixed effects alone was relatively high, marginal $R^2 = .40$ (Nakagawa & Schielzeth, 2013).

[INSERT TABLE 6 HERE]

Asian Participants

The coefficients for the model fit to the data from Asian participants are displayed in Table 7. Again, only two predictors were significant – the FMT, which here estimated the difference between the Asian (this time coded as zero) and the Australia CFMTs, b = -5.86, $t(420) = 15.23$, $p < .001$, demonstrating a significant OEE effect. This means that for Asian participants, scores on the Australia CFMT were on average 5.85 points lower than for the Asia CFMT. Additionally, there was a significant coefficient for the Boston CFMT score, which here represented a measure of independent other-ethnicity performance, $b = 5.31$, $t(757.82) = 15.83$, $p < .001$. Here, this represents the pattern that a one standard deviation increase in *other*-ethnicity recognition performance is associated with, on average, a change of 5.31 units in own-ethnicity performance as measured by the Asia CFMT. The lack of significant interaction between the FMT and Boston predictor here (b = -0.16) indicates that this relationship is practically equivalent between the Boston and the Australia CFMT scores. The variance explained in the Australia and Asia CFMT scores was as similarly high as the model built on Caucasian data, marginal $R^2 = .45$.

[INSERT TABLE 7 HERE]

Examining model predictions

The estimated statistical models thus far demonstrate significant OEEs, and an influence of the Boston CFMT scores on the Australia and Asian CFMTs, whether that represents an own- or other-ethnicity measure of recognition performance. Examining the predictions made by the models is key to their interpretation. As the models essentially fit a separate slope for the Australia and Asia CFMTs simultaneously (coded by the FMT coefficient), and by allowing these separate slopes to interact with the other predictors, we are

able to examine the likely OEE at high and low levels of the Boston CFMT and social contact scores. Figure 3 demonstrates the predictions of each model, derived by using the models to predict scores separately for the Australia and Asia CFMTs for hypothetical participants with varying scores on the Boston and social contact measures. The figure makes it clear that the OEE – the difference between the slopes of the Asia and Australia CFMTs – is consistent at all various combinations of low and high Boston and social contact measures, evaluated here at scores ranging from ±2SDs on the predictors. Indeed, this consistency is clear from the lack of interactions in the model. These predictions thus allow us to examine how individuals with excellent or very poor own- or other-ethnicity performance and high or low levels of contact might do on tests of own- or other-ethnicity performance, but with information estimated from a full range of data as opposed to smaller samples.

[INSERT FIGURE 3 HERE]

Further considerations and robustness checks

An additional possible source of variability we have not considered so far is that participants were sampled from different countries within our models – that is, not all Caucasian and Asian participants were from the same countries, as described in the method. It is therefore possible that variation within those countries in terms of face processing ability or otherwise could have an impact on our results.

To test this, we recreated our two models, but this time included an additional random intercept for country alongside that of participants representing their country of origin (i.e. whether Asian participants were from South Korea, China, Japan, or Singapore, and Caucasian participants were from the UK, Australia, or Serbia). Treating country of origin as a random factor is appropriate as we wish to make inferences about countries that are

generally Asian or Caucasian, and our data represents only a sample of the possible countries that fit this profile. We compared these new models to the original models used in the analyses without the additional random intercept using a likelihood ratio test, to confirm whether the more complex model had a better fit to the data. For both the Caucasian model, the likelihood ratio test was not significant - $\chi^2(1) = 0.00$, $p = .999$. For the Asian model, this test was significant, $\chi^2(1) = 6.97$, $p = .008$ – indicating that country of origin did improve the fit to the data. Examining the AIC of the model showed a small change between models (without = 5511.5, with = 5506.5), and the marginal R^2 of the model increased by 0.01%, from 0.446 to 0.456. The overall pattern of results were unchanged.

We also estimated our models by swapping the positions of the Australia and Boston CFMTs, by using Australia scores as the independent measure of performance and Boston scores being predicted alongside the Asian CFMT. No differences in the overall conclusions were found. We also sought to examine the stability of the OEE effect by using random split-half resampling techniques, which showed the magnitude of the OEE was very consistent. See the Supplementary Materials for details.

General Discussion

The current study represents the largest ever undertaken investigating the OEE across within population distributions of own-ethnicity and other-ethnicity face recognition performance. Our results demonstrated the following key findings:

1. Our study finds a robust OEE effect in both Asian and Caucasian samples, replicating previous studies of the OEE using the CFMT paradigm.
2. Our modelling approach allowed us to test whether the magnitude of the OEE varied in relation to individual levels of own ethnicity OR other ethnicity ability. It did not. Our model therefore shows a remarkably consistent impact of the OEE across the entire range of the populations investigated.
3. Our approach also allows us to test whether social contact impacts the OEE – and we found no evidence for this. But, it is of note that in our case the range of scores on our measure of social contact was not substantial (with contact scores being relatively low), despite the fact that we sampled across a number of different countries. In any case, a meta-analysis of OEE research articles demonstrated that self-report assessments of other-ethnicity contact explained less than 3% of the total variance in the OEE (Meissner & Brigham, 2001), indicating that factors beyond the kind of measures we have implemented on this issue may be more key to modulating the OEE in individual performance (e.g., such as bilingualism; Burns et al., 2019).
4. Our model also indicates no combination of these factors appear to impact scores on CFMTs of own or other ethnicity (i.e., no evidence of a two or three-way interaction).

In summary, this work demonstrates that an OEE is a consistent feature of face recognition performance for participants sampled across a variety of nations and cultures – and in addition this differentiation in performance, which could be characterised as either an own-ethnicity advantage or an other-ethnicity disadvantage, is consistent in magnitude across all individuals. Our finding that individuals at 'extremes' of own ethnicity performance show an equivalent OEE is consistent with previous work undertaken with groups of individuals classified as *developmental prosopagnosia* (DeGutis, et al., 2011; Cenac et al., 2019) and *super-recognisers* (Bate et al., 2018; Robertson et al., 2019). In that in both cases, the evidence emerging from testing of such populations suggests both groups show OEEs; our work builds on this by further indicating that the quality of this OEE is indeed no different from that of individuals at any other points on the distribution of own-ethnicity recognition ability.

However, our findings, at least initially, may be seen to be contrary to those of Wan et al., (2017) and their reports of individuals with putative *other-ethnicity blindness*, in that we found no evidence that the quality of the OEE differed even with individuals who performed at the lowest end of the distribution of other-ethnicity face recognition accuracy. It should be noted that an advantage of our work is that by considering this issue across the full distribution of population performance, we avoided issues around classification 'cut-off' (i.e., 2 SDs) discussed earlier. In the Wan et al. (2017) study, poor performers were selected on the basis of a somewhat arbitrary statistical distinction, albeit an approach often used by others – and this classification was not confirmed with any further testing. Thus, it remains possible that in their work, the observed differences between two tests could be simply due to the fact that the same test was used as the classifier *and* as the comparator. It is therefore likely that these differences in our approaches may explain the potentially contrary findings.

However, it should be noted that social contact was quite limited in variability in all our participant cohorts – and we suggest this may explain why no effect of social contact was seen despite previous reports indicating an influence (e.g., Zhou et al., 2019). This key issue may also explain our initially contrary findings to Wan et al. (2017); where it is possible that if across a test group there is considerable variability in social contact, a small sub-group may have much lower social contact than the rest of the group in general. If so, that sub-group might perform much worse relative to the rest of the group, and thus reach the classification of 2SDs below the mean. Importantly, Wan et al., (2017) reported that of the 37 participants who met criteria for being very poor with other-ethnicity faces (i.e., 2SDs below mean accuracy), 36/37 had reported low contact with individuals of the other ethnicity, and thus it's likely what is driving the presence of very poor other-ethnicity face accuracy is a process linked to social contact rather than face processing in general. Given the low variability of contact in our samples, we would suggest this could explain why we found no evidence of any individuals with a relative *other-ethnicity blindness.* We therefore agree with the conclusions of Wan et al. (2017) that the presence of individuals who would meet such a criteria is likely dependent on the relative individual variability of social contact for the group tested and not to do with the base level of face recognition performance generally. As a consequence, we are reluctant to draw the more general conclusion that social contact does not influence the magnitude of OEE and it would be interesting to test this in a sample with a much more varied pattern of social contact than we were able to obtain.

Furthermore, it is important to stress that our modern society allows for more varied types of social contact than the face-to-face interactions that traditionally defined 'social contact', as measured by the questionnaire used in the current study. For example, East-Asian pop bands have been increasing in popularity in the Western media through films, music videos, and advertisements, among others, and vice versa. This type of cultural contact is not

covered in the contact measure that was used in this study, but could potentially have a considerable impact on individuals' ability to recognise and discriminate between faces of other-ethnicities – simply because they can provide many more opportunities to increase exposure to faces from other ethnicities beyond contact in the traditional sense. We would therefore suggest that this needs to be incorporated in future studies that aim to measure contact with other ethnic groups.

With this in mind, consider the case of individuals who appear to perform very poorly with own-ethnicity faces; what is striking from our work is that *despite* the issues with own-ethnicity faces such individuals have, they manifest an OEE commensurate with the rest of the population. This clearly indicates that whatever unpins the challenges faced by such individuals with faces of their own ethnicity, this is independent of the OEE. We would speculate that this reflects the fact that all individuals, independent of natural face recognition ability, can still gain *some* visual learnt experience from own-ethnicity faces. This learnt experience underpins a remaining advantage for own-ethnicity faces (or disadvantage for other-ethnicity faces) and hence an OEE is consistently present. We therefore interpret our findings in a similar manner to that of Cenac et al. (2019) – namely, that face processing is underpinned by two key factors: on the one hand there is a form of *inherited susceptibility* to generally poor face processing ability and on the other hand there is a *visual learnt experience* factor that can drive differential performance across types of face ethnicity. What our work clearly demonstrates is that variability on the first of these factors has no impact on the magnitude of the OEE in face recognition memory – regardless of an inherited susceptibility to being generally poor or very good with faces, all other things being equal, there is always a consistent and universal 'fixed' benefit/cost to recognition memory across faces of differing ethnicities. The consistent nature of this OEE effect also implies that if inherited susceptibility to generally very good face processing ability is the case for a given

individual, although that person will perform more poorly with other-ethnicity faces, they will still be largely superior to all other individuals in that same population. Thus making the case that in practical terms, the best persons to employ for passport control will always be superior face recognisers in a given population.

Earlier in the introduction we mentioned that previous work has demonstrated that the OEE can be moderated through participant training with other-ethnicity faces (e.g., Lebrecht et al., 2009). An account for this training effect has been linked to the suggestion that differential performance across face ethnicities may be underpinned by the degree of configural or featural processing being used. That is, it is likely that own-ethnicity faces, given their high degree of familiarity, implicate a different 'bias' toward configural/holistic versus featural/part-based processing, (Hayward et al., 2008; Zhao et al., 2014; Rhodes et al., 2010). With such an explanation in mind, DeGutis et al. (2011) has suggested that training can mediate attentional 'bias' across own/other ethnicity faces such that it 'boosts' configural processing of other ethnicity faces. Although we did not examine the issue of configural/featural processing, we might speculate that the consistent and 'fixed' OEE pattern we see across all levels of individual ability in our work, is the consequence of this consistent attentional 'bias' across faces of different types.

This raises an interesting future avenue of research regarding the effects of training on the OEE – previous studies have largely considered such effects at the group level (e.g., Tanaka & Pierce, 2009), and we suggest rather taking an individual differences approach – thus exploring the consequences of training across individual variability in own ethnicity face recognition. For example, although the work by DeGutis et al., (2011) speaks to the question of the impact of training for individuals at the lowest end of performance (i.e., developmental prosopagnosia), it would be interesting to explore the consequences of training across all levels of individual ability using a similar approach to that undertaken here. If the OEE

reflects a fixed ‘cost’ of a strategic ‘bias’ in attentional resource allocation for configural processing across faces of different ethnicities, and training can reduce this ‘bias’, the prediction should be that all levels of ability would see the same relative reduction in OEE magnitude. Put simply, if the OEE reflects the consistent impact of a strategic attentional ‘bias’, then it should be possible via training for all types of faces to reach optimal performance commensurate with own-ethnicity face testing for a given individual.

A final consideration is the issue of statistical power and potential measurement error. Our large sample and use of linear mixed models afford greater power, and our power analysis (see SM) indicated that we can comfortably detect changes in CFMT scores as small as .20 - .25 across our predictors. Notably, some of the coefficients in our two models were estimated to be below this threshold, and as such, we cannot explicitly rule out the absence of an effect here (i.e., there may be an interaction between the OEE and own-ethnicity performance) that is too small to detect with our sample size. An important factor that may contribute to increasing ‘noise’ in our pattern of results is that the OEE itself has low measurement reliability. As a means of mitigating the potential contribution of poor measurement, we selected three face recognition measures with well-established reliability (see Horry et al., 2015; McKone et al., 2012), and this was also confirmed in our own analyses. However, just because an individual measure is reliable, does not therefore entail that the *product* of two such measures (that is the difference between the two, which is how the OEE is defined) is necessarily also reliable (see Ross et al., 2014; Sunday et al., 2017). As a consequence we undertook reliability analyses on our OEE effect, and report that indeed reliability levels are lower than is seen for the individual tests themselves, but still sufficiently high for us to have some confidence in our interpretation of the lack of interactions seen in our analyses. This is in fact the first time such reliability analyses have been undertaken and they provide an important caveat to the findings of OEE studies both

past, present and future – since if one assumes that our pattern is often the case (that the OEE is less reliable than the individual tests from which it is computed), the individual tests used *must* be very high in reliability in the first place and it would be good practice for OEE reliability to be reported if not.

A final point is that, for the interaction terms that did not reach statistical significance, the coefficient estimates were very small. Since estimates of coefficients using least squares are unbiased (i.e., on average, the coefficients will represent the effect in the population), and our sample is large enough to provide a stable estimate, we would tentatively conclude that any interaction terms between the OEE and other factors are likely to be small in practical terms. For example, for Caucasian participants, the three-way interaction coefficient was .24 units, which is much less than a single unit on a given CFMT, and thus unlikely to translate into a qualitative 'real-world' difference in recognizing faces of another ethnicity. However, we also recognize that the issue of measurement error is at play here, and this difference could be larger than this. We did however build our statistical models for our analysis to mitigate these limitations, as the inclusion of the random intercept term means the fixed effect of the OEE is scaffolded by individual level intercepts, and therefore we are confident that such issues were minimal for our current data.

In sum, the current work is the first to consider the OEE from the perspective of individual variability across a variety of nations and cultures; our message is that the magnitude of the OEE is of a consistent quality across all levels of ability seen both from the perspective of variance on own ethnicity face recognition performance and other ethnicity face recognition performance. These findings are consistent with studies that have focused their attention on sub-groups of individuals at both the bottom (i.e., developmental prosopagnosia, DeGutis et al., 2011) and top (i.e., super-face recognisers, Bate et al., 2018) of

the population distribution, in that OEE patterns were also reported in their samples – our work builds on this by demonstrating such effects are by no means qualitatively different. Intriguingly, given the OEE we found across individuals was consistent in magnitude, we speculate that this is compatible with an attentional 'bias' account for the OEE (as suggested by DeGutis et al., 2011) – essentially, the OEE reflects the utilisation of a face-based strategic attentional processing 'bias', which incurs a benefit/cost to recognition memory across own/other ethnicity faces. This impact is independent of the general level of face recognition memory for any given individual and thus the OEE remains of consistent magnitude across all levels of ability. It would be interesting for future work to explore this issue further, perhaps by considering the impact of training through the lens of individual variability.

References

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., Tanaka, J. W., & Lee, K. (2013). Developmental origins of the other-race effect. *Current Directions in Psychological Science*, *22*(3), 173-178. https://doi.org/10.1177/0963721412474459

Barrera, M. E., & Maurer, D. (1981). Recognition of mother's photographed face by the three-month-old infant. *Child development*, 714-716. https://www.jstor.org/stable/pdf/1129196.pdf

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. https://doi.org/10.1080/17470218.2016.1195414

Bate, S., Bennetts, R., Hasshim, N., Portch, E., Murray, E., Burns, E., & Dudfield, G. (2018). The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 363-377. https://doi.org/10.1037/xhp0000607

Bate, S., Bennetts, R.J., Tree, J.J., Adams, A., Murray, E. (2019). The domain-specificity of face matching impairments in 40 cases of developmental prosopagnosia. *Cognition, 192,* 104031. https://doi.org/10.1016/j.cognition.2019.104031

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. doi: 10.18637/jss.v067.i01

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., ... & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423-455.

Burns, E. J., Bennetts, R. J., Bate, S., Wright, V. C., Weidemann, C. T., & Tree, J. J. (2017). Intact word processing in developmental prosopagnosia. *Scientific reports*, *7*(1), 1-12. https://doi.org/10.1038/s41598-017-01917-8

Burns, E. J., Martin, J., Chan, A. H., & Xu, H. (2017). Impaired processing of facial happiness, with or without awareness, in developmental prosopagnosia. *Neuropsychologia*, *102*, 217-228. https://doi.org/10.1016/j.neuropsychologia.2017.06.020

Burns, E. J., Tree, J., Chan, A. H., & Xu, H. (2019). Bilingualism shapes the other race effect. *Vision research*, *157*, 192-201. https://doi.org/10.1016/j.visres.2018.07.004

Burns, E.J., Tree, J.J., Weidemann, C.T. (2014). Recognition memory in developmental prosopagnosia: electrophysiological evidence for abnormal routes to face recognition *Frontiers in human neuroscience 8*, 622. https://doi.org/10.3389/fnhum.2014.00622

Cenac, Z., Biotti, F., Gray, K. L., & Cook, R. (2019). Does developmental prosopagnosia impair identification of other-ethnicity faces?. *Cortex*, *119*, 12-19. https://doi.org/10.1016/j.cortex.2019.04.007

Cho, S. J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., ... & Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological assessment*, *27*(2), 552.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 879-894. https://doi.org/10.1080/14640749508401421

Crookes, K., & Rhodes, G. (2017). Poor recognition of other-race faces cannot always be explained by a lack of effort. *Visual Cognition*, *25*(4-6), 430-441. https://doi.org/10.1080/13506285.2017.1311974

DeGutis, J., DeNicola, C., Zink, T., McGlinchey, R., & Milberg, W. (2011). Training with own-race faces can improve processing of other-race faces: Evidence from

developmental prosopagnosia. *Neuropsychologia*, *49*(9), 2505-2513. https://doi.org/10.1016/j.neuropsychologia.2011.04.031

DeGutis, J., Mercado, R. J., Wilmer, J., & Rosenblatt, A. (2013). Individual differences in holistic processing predict the own-race advantage in recognition memory. *PloS one*, *8*(4), e58253. doi: 10.1371/journal.pone.0058253

De Heering, A., De Liedekerke, C., Deboni, M., & Rossion, B. (2010). The role of experience during childhood in shaping the other-race effect. Developmental science, 13(1), 181-187.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576-585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Estudillo, A. J., Lee, J. K. W., Mennie, N., & Burns, E. (2020). No evidence of other-race effect for Chinese faces in Malaysian non-Chinese population. *Applied Cognitive Psychology*, *34*(1), 270-276. https://doi.org/10.1002/acp.3609

Freckleton, R. P. (2002). On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *Journal of Animal Ecology*, 71(3) 542-545. **https://doi.org/10.1046/j.1365-2656.2002.00618.x**

Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *Neuroimage*, *22*(4), 1628-1635. https://doi.org/10.1016/j.neuroimage.2004.03.049

Goldstein, A. G., & Chance, J. E. (1985). Effects of training on Japanese face recognition: Reduction of the other-race effect. *Bulletin of the Psychonomic Society*, *23*(3), 211-214. https://link.springer.com/content/pdf/10.3758/BF03329829.pdf

Hancock, K. J., & Rhodes, G. (2008). Contact, configural coding and the other-race effect in face recognition. *British Journal of Psychology*, *99*(1), 45-56. https://doi.org/10.1348/000712607X199981

Harvey, A. J. (2014). Some effects of alcohol and eye movements on cross-race face learning. Memory, 22(8), 1126-1138.

Hayward, W. G., Rhodes, G., & Schwaninger, A. (2008). An own-race advantage for components as well as configurations in face recognition. *Cognition*, *106*(2), 1017-1027. https://doi.org/10.1016/j.cognition.2007.04.002

Horry, R., Cheong, W., & Brewer, N. (2015). The other-race effect in perception and recognition: insights from the complete composite task. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 508-524. http://dx.doi.org/10.1037/xhp0000042

Jackson, M. C. Counter, P. & Tree, J. J (2017). Intact short-term memory for faces in developmental prosopagnosia. Neuropsychologia, 106, 60-70. https://doi.org/10.1016/j.neuropsychologia.2017.09.003

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, *18*(12), 1084-1089. https://doi.org/10.1111/j.1467-9280.2007.02029.x

Lebrecht, S., Pierce, L. J., Tarr, M. J., & Tanaka, J. W. (2009). Perceptual other-race training reduces implicit racial bias. *PloS one*, *4*(1), e4215. doi: 10.1371/journal.pone.0004215

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of personality and social psychology*, *13*(4), 330-334. http://dx.doi.org/10.1037/h0028434

MacLin, O. H., Van Sickler, B. R., MacLin, M. K., & Li, A. (2004). A re-examination of the cross-race effect: The role of race, inversion, and basketball trivia. North American Journal of Psychology, 6(2), 189-204.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R. B., Rivolta, D., ... & O'Connor, K. B. (2011). Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test–Australian. *Cognitive neuropsychology*, *28*(2), 109-146. https://doi.org/10.1080/02643294.2011.616880

McKone, E., Stokes, S., Liu, J., Cohan, S., Fiorentini, C., Pidcock, M., ... & Pelleg, M. (2012). A robust method of measuring other-race and other-ethnicity effects: The Cambridge Face Memory Test format. *PLoS One*, *7*(10), e47956. doi: 10.1371/journal.pone.0047956

McKone, E., Wan, L., Robbins, R., Crookes, K., & Liu, J. (2017). Diagnosing prosopagnosia in East Asian individuals: Norms for the Cambridge Face Memory Test–Chinese. *Cognitive neuropsychology*, *34*(5), 253-268. https://doi.org/10.1080/02643294.2017.1371682

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy & Law*, 7, 3–35. http://dx.doi.org/10.1037/1076-8971.7.1.3

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, *4*(2), 133-142. DOI: 10.1111/j.2041-210x.2012.00261.x

Ng, W. J., & Lindsay, R. C. (1994). Cross-race facial recognition: Failure of the contact hypothesis. *Journal of Cross-Cultural Psychology*, *25*(2), 217-232. https://doi.org/10.1177/0022022194252004

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology, 110*(3), 461-479. https://doi.org/10.1111/bjop.12368

Rhodes, G., Ewing, L., Hayward, W. G., Maurer, D., Mondloch, C. J., & Tanaka, J. W. (2010). Contact and other-race effects in configural and component processing of faces. *British Journal of Psychology*, *100*(4), 717-728. https://doi.org/10.1348/000712608X396503

Robertson, D. J., Black, J., Chamberlain, B., Megreya, A. M., & Davis, J. P. (2019). Super-recognisers show an advantage for other race face identification. *Applied Cognitive Psychology*, *34*(1) https://doi.org/10.1002/acp.3608

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior research methods*, *47*(3), 736-743.

Sangrigoli, S., & De Schonen, S. (2004). Recognition of own-race and other-race faces by three-month-old infants. *Journal of Child Psychology and Psychiatry, 45*(7), 1219-1227. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7610.2004.00319.x

Sunday, M. A., Richler, J. J., & Gauthier, I. (2017). Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Attention, Perception, & Psychophysics*, *79*(5), 1453-1465.

Tanaka, J. W., & Pierce, L. (2009). The neural plasticity of other-race face recognition. *Cognitive, Affective, & Behavioral Neuroscience*, *9,* 122–131. https://link.springer.com/content/pdf/10.3758/CABN.9.1.122.pdf

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), 1-9. http://web.uvic.ca/psyc/vizcoglab/pubPDFs/cognitionproofs.pdf

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, 679-686.

Walker, P. M., & Hewstone, M. (2006). A perceptual discrimination investigation of the own-race effect and intergroup experience. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *20*(4), 461-475. https://doi.org/10.1002/acp.1191

Wan, L., Crookes, K., Reynolds, K. J., Irons, J. L., & McKone, E. (2015). A cultural setting where the other-race effect on face recognition has no social–motivational component and derives entirely from lifetime perceptual experience. *Cognition*, *144*, 91-115. https://doi.org/10.1016/j.cognition.2015.07.011

Wan, L., Crookes, K., Dawel, A., Pidcock, M., Hall, A., & McKone, E. (2017). Face-blind for other-race faces: Individual differences in other-race recognition impairments. *Journal of Experimental Psychology: General*, *146*(1), 102-122. https://doi.org/10.1037/xge0000249

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of sciences*, *107*(11), 5238-5241.

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225-230. https://journals.sagepub.com/doi/pdf/10.1177/0963721417710693

Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2003). Inter-racial contact and the own-race bias for face recognition in South Africa and England. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *17*(3), 365-373. https://doi.org/10.1002/acp.898

World Medical Association. (2009). Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Jahrbuch Für Wissenschaft Und Ethik*, *14*(1), 233-238. https://doi.org/10.1515/9783110208856.233

Zhao, M., Hayward, W. G., & Bülthoff, I. (2014). Holistic processing, contact, and the other-race effect in face recognition. *Vision Research*, *105*, 61-69. https://doi.org/10.1016/j.visres.2014.09.006

Zhou, X., Elshiekh, A., & Moulson, M. C. (2019). Lifetime perceptual experience shapes face memory for own-and other-race faces. *Visual Cognition*, 1-14. https://doi.org/10.1080/13506285.2019.1638478

Table 1

Summary of key studies' findings relating to the other-ethnicity effect in relation to geographical and self-report contact

Study	Test comparisons	Samples	OEE
Chiroro & Valentine, 1995	Old-new	Africans and Caucasians living in Harare, Zimbabwe (high contact) Caucasians in UK (low contact) Africans in South Zimbabwe (low contact)	Hits: High contact group had similar levels of hits for both African and Caucasian faces compared to low contact groups. False Positives: High contact Africans had lower FP compared to the other groups.
De Heering et al., 2010 (Geographical)	Old-new	Adopted Asian children in Belgium and Caucasian children	Caucasian children showed OEE; Asian children showed similar recognition of Asian and Caucasian faces.
Hancock & Rhodes, 2008 (Self-report)	Recognition tasks using upright and inverted images	Chinese and Caucasians living in Australia (varied arrival times)	Increased contact with the other-ethnicity predicted lower OEE in recognition of upright faces, and reduced inversion effects.
Harvey, 2014 (Self-report)	Old-new	Caucasian students tested on Caucasian and Indian faces	No significant effect of contact levels on recognition performance of Indian faces.
MacLin et al., 2004 (Self-report)	Recognition tasks using upright and inverted images	Caucasian students who were categorised as either Novices/Experts in African-American basketball players	No inversion effects found in both groups.

Ng & Lindsay, 1994 (Self-report)	Old-new	Study 1: Caucasians and Asians living in Canada (Asians reported *high* contact with Caucasians) Study 2: Caucasians and Asians living in Singapore (Caucasians reported *low* contact with Asians)	Study 1: Asians showed similar FA rates for both Caucasian and Asian faces. Self-report contact was not significantly related to recognition performance. Study 2: Caucasians recognized both types of faces equally. Caucasians in Singapore did not have a significantly different recognition performance compared to Caucasians in Canada.
Rhodes et al., 2010 (Geographical and self-report)	Caucasian and Asian faces - blurred faces and scrambled faces	Chinese students living in Australia (varied in arrival time)	Hits and false alarm rates (d') had a negative correlation with duration of stay in Australia. Self-report contact did not reach significance.
Tanaka et al., 2004 (Geographical and self-report)	Part-whole task	Caucasians and Asians living in Germany (Asians reported *high* contact with Caucasians)	Caucasians = high recognition of whole face for Caucasian faces, low recognition of whole and part faces for Asian faces. Asians = no significant difference in the recognition of part or whole faces for both face types.
Wright et al., 2003 (Geographical and self-report)	Old-new	Blacks and Whites living in South Africa (high contact) Whites in UK (low contact)	Hits and false alarm rates (d') of Black African population were significantly negatively correlated with self-report contact.
Zhou et al., 2019 (Geographical and self-report)	Cambridge Face Memory Test (CFMT) Australia and Chinese	Chinese individuals living in Australia (varied arrival time) Caucasians	Higher contact (longer time spent in Toronto and higher self-report contact) with Caucasians

Zhao et al., 2014 (Self-report)	Part-whole task, blurred and scrambled task, CFMTs	Chinese and Germans	Higher contact predicted smaller OEE in CFMTs, whole condition, and blurred condition, compared to part and scrambled conditions.

Note: Studies which used facial recognition or facial perception tests and measured amount of contact and other-ethnicity effect.

Table 2

Summary of key studies' findings relating to the other ethnicity effect using CFMT

Study	Test comparisons	Samples	OEE
Zhou et al., 2019	Boston-Asian	Caucasian	d = .64
DeGutis et al., 2013	Boston - Asian	Caucasian	d = .5
Crookes & Rhodes, 2017	Australian - Asian	Caucasian	d = 1.04 (standard) d = 1.24 (self-paced)
Horry et al., 2015	Australian - Asian	Caucasian	d = .91
		Asian	d = 1.14
Wan et al., 2015	Australian - Asian	Caucasian	% difference = 7.25
		Asian	% difference = 8.84
McKone et al., 2012	Australian - Asian	Caucasian	d = .76
		Asian	d = .84

Note: List of studies which used two versions of CFMT (own- versus other-race) to measure OEE. Note that all studies reported a robust effect, thus implying that in normative population, individuals are better at recognizing faces from their own- compared to those from other-ethnicities.

Table 3

Participant count, age means and standard deviations in the sample

Sample	Country	Age Mean	Age SD	Female, Male	Total sample
Caucasian	Australia	19.54	1.99	71, 31	102
	Britain	18.67	0.93	159, 36	195
	Serbia	20.26	1.49	56, 47	103
Asian	China	19.05	0.95	61, 42	103
	Japan	19.77	1.58	62, 58	120
	South Korea	20.37	1.18	53, 56	109
	Singapore	20.49	1.33	68, 24	92
Grand		19.61	1.51	530, 294	824

Note: Descriptive statistics of the sample cohort shown for each country, ethnic group, and grand total.

Table 4

Mean scores and standard deviations of contact scores

	Know		Social		Individuation		Mean Contact	
Country	M	SD	M	SD	M	SD	M	SD
Australia	1.66	0.97	1.91	1.02	2.34	1.16	1.97	0.83
Britain	1.5	0.76	1.56	0.86	1.85	0.99	1.63	0.67
China	1.07	0.25	1.38	0.81	2.13	0.35	1.53	0.33
Japan	1.51	0.84	1.16	0.4	1.24	0.53	1.3	0.46
Korea	1.85	1.5	1.28	0.57	1.47	0.76	1.53	0.61
Serbia	1.05	0.26	1.12	0.44	1.16	0.49	1.11	0.3
Singapore	1.3	0.72	1.3	0.47	1.7	0.76	1.43	0.45

Note: The means for the number of people known (Q1), social (Q2-5) and individuation (Q6-10) components of the Social Contact Scale (SCS, Walker & Hewstone, 2006) used in this study did not show significant variance, allowing the authors to collapse the scores to create a composite contact measure which was used in the subsequent analyses. The original scores for social and individuation components of the SCS were inversed, i.e. higher scores mean lower contact, however, for the linear model analysis, we needed the scores across all variables to be in the same direction, e.g. higher scores mean better recognition skills and higher contact with other-ethnicity group. Therefore, items 2-10 in the SCS were reverse scored, and the three scales were averaged together to create a Mean Contact score where higher scores reflect higher contact.

Table 5

Descriptive statistics for the country cohort on CFMT measures.

Country		Asian		Australian		Boston	
	N	Mean	SD	Mean	SD	Mean	SD
Australia	102	50.9	7.88	55.15	7.47	55.94	7.87
Serbia	103	51.04	8.22	57.69	7.35	58.14	8.61
UK	195	52.2	8.61	54.37	7.66	55.19	8.45
Overall Caucasian	400	51.57	8.33	55.42	7.64	56.14	8.41
China	103	56.59	8.31	48.73	7.52	47.32	8.9
Japan	120	56.73	7.5	48.78	7.41	51.82	7.43
South Korea	109	55.5	8.98	52.72	8.29	51.44	8.01
Singapore	93	55.11	7.55	50.65	8	49.08	8
Overall Asian	424	56.03	8.11	50.19	7.95	50.04	8.26
Total	824	53.87	8.51	52.73	8.22	53	8.87

Note: Mean correct scores (over 72 items; chance performance is ≤ 24) and standard deviations for the three CFMT versions used in this study for each country cohort and ethnic groups.

Table 6

Parameter estimates for the Caucasian participants' model

Parameter	*b* [95% CI]	*SE*	*t*-value	*p*-value
Intercept	53.64 [52.97, 54.31]	0.34	157.09	< .001
FMT				< .001
(0 = Australia)	-3.98 [-4.77, -3.18]	0.41	-9.76	
Boston	5.03 [4.37, 5.69]	0.34	14.98	< .001
Contact	0.03 [-0.53, 0.6]	0.29	0.11	0.911
Boston * FMT	0.16 [-0.63, 0.94]	0.4	0.39	0.698
Contact * FMT	0.55 [-0.12, 1.23]	0.34	1.61	0.107
Boston * Contact	-0.14 [-0.73, 0.46]	0.31	-0.45	0.654
FMT * Boston * Contact	0.24 [-0.48, 0.95]	0.36	0.65	0.516

Note: Estimates for the Caucasian model showing FMT scores significantly influence the variability in the scores. Boston-CFMT scores was used as own-ethnicity measure. Contact scores do not show significant contribution in the FMT scores, indicating that level of contact in this study do not influence other-ethnicity face recognition.

Table 7

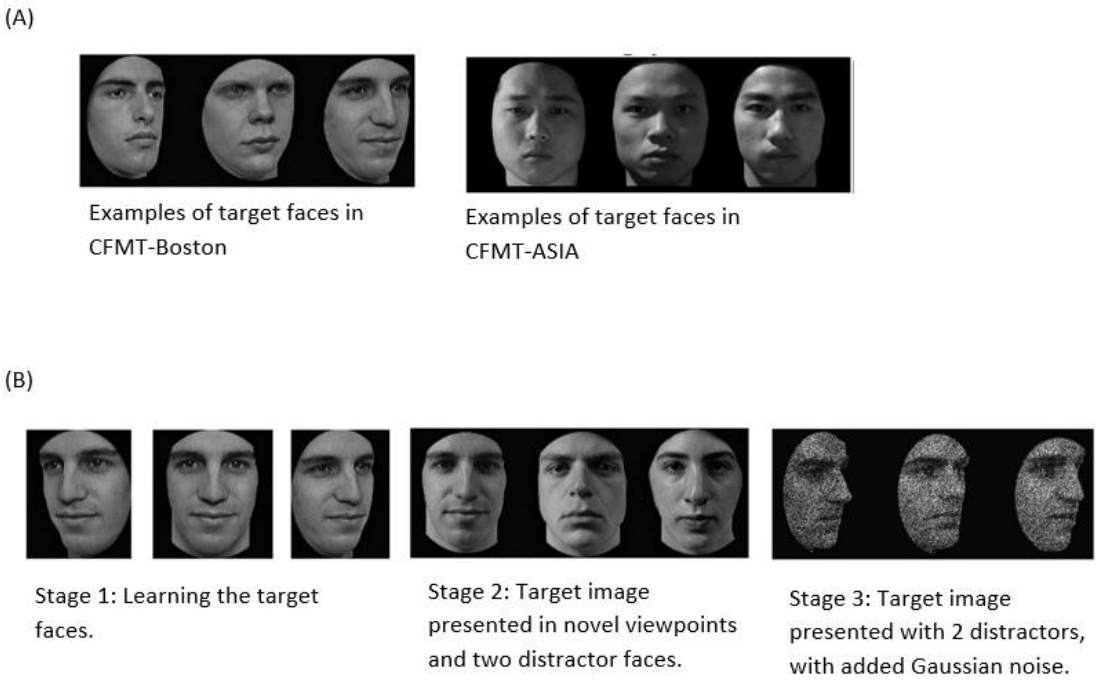
Parameter estimates for the Asian participants' model

Parameter	*b* [95% CI]	*SE*	*t*-value	*p*-value
Intercept	57.77 [57.11, 58.42]	0.33	173.91	< .001
FMT				< .001
(0 = Asia)	-5.86 [-6.61, -5.1]	0.38	-15.23	
Boston	5.31 [4.65, 5.97]	0.34	15.77	< .001
Contact	-0.4 [-1.2, 0.4]	0.41	-0.97	0.331
FMT * Boston	-0.17 [-0.93, 0.6]	0.39	-0.42	0.671
FMT * Contact	0.23 [-0.7, 1.16]	0.47	0.49	0.622
Boston * Contact	-0.1 [-0.96, 0.76]	0.44	-0.23	0.822
FMT * Boston * Contact	-0.44 [-1.44, 0.55]	0.51	-0.87	0.382

Note: Estimates for the Asian model showing FMT scores significantly influence the variability in facial recognition scores. Boston-CFMT scores were used as other-ethnicity measure. Similar to the Caucasian model, contact scores do not show significant contribution in the FMT scores, indicating that level of contact in this study does not influence other-ethnicity face recognition.

Figure 1

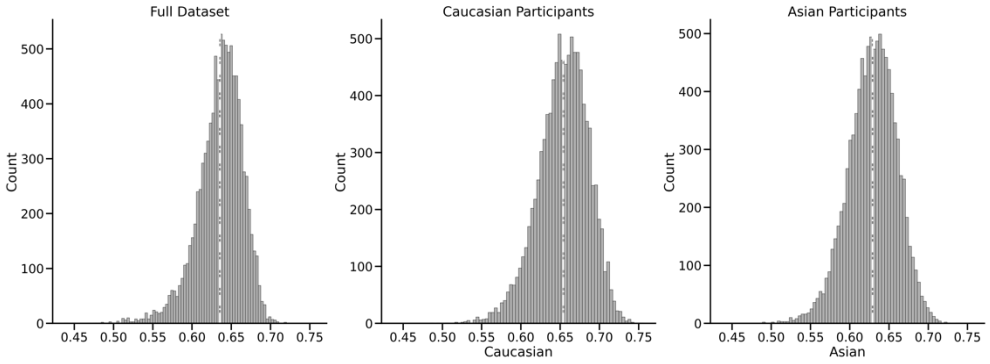
Phases of Cambridge Face Memory Test



Note: (A) Examples of target faces in CFMT-Boston (Duchaine & Nakayama, 2006) and CFMT-Asia (McKone et al, 2012). (B) Illustrative images for all CFMT procedures. For full details of the procedures, see Duchaine and Nakayama (2006).

Figure 2

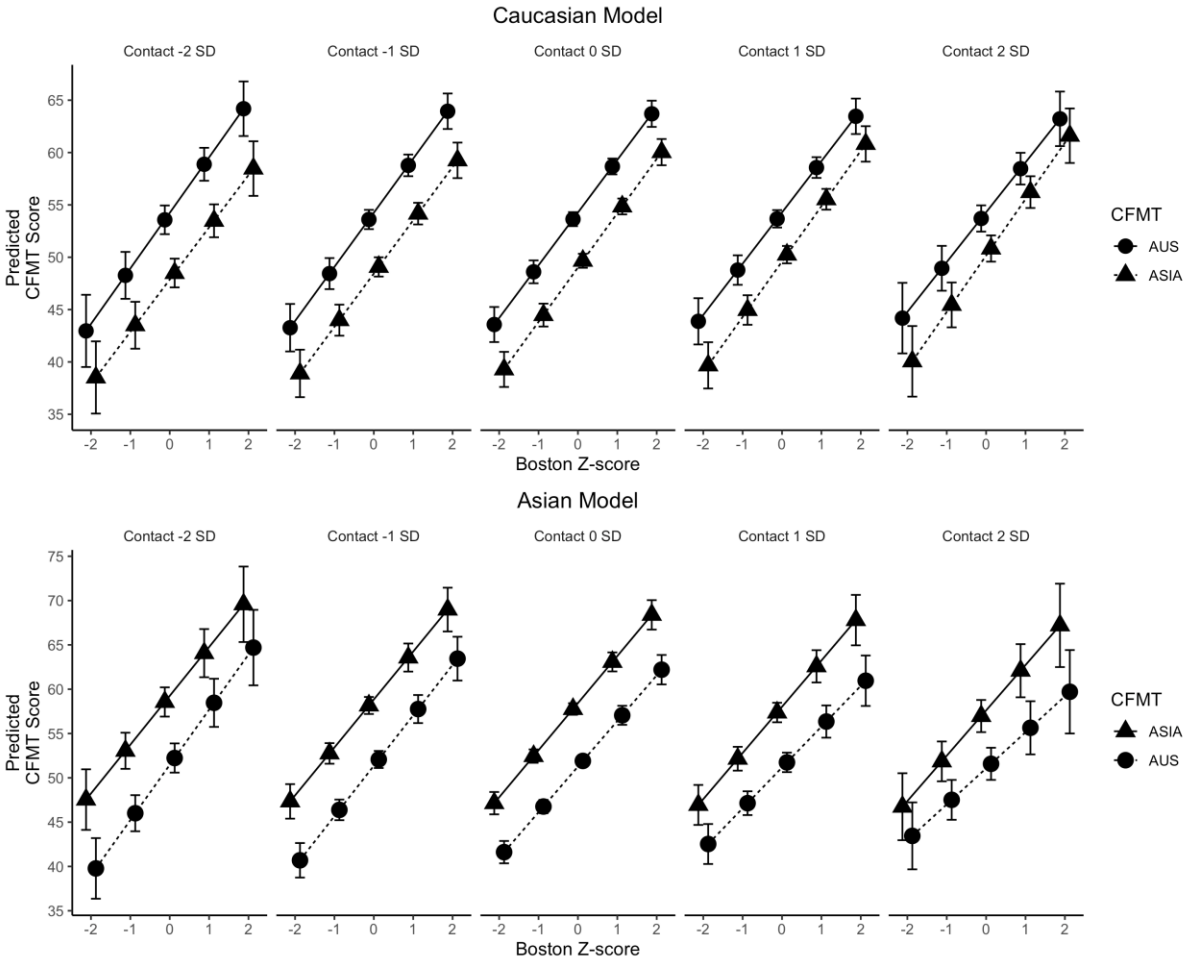
Reliability analysis for OEE scores



Note: Distributions of the reliability of the OEE generated by bootstrap resampling. The average of each distribution is marked by the dashed white line.

Figure 3

OEE magnitude in Caucasian (top) and Asian (bottom) participant models



Note: Predictions of the Caucasian participants model top, by varying levels of contact (separate axes) and the Boston CFMT (X-axis). Error bars represent 95% confidence intervals of the FMT scores. For the top axis, the Boston represents an independent measure of own-ethnicity performance, and for the bottom axis, it represents a measure of other-ethnicity performance.