Running Head: Prevalence of individuals with "Other Ethnicity Blindness"

How prevalent is "Other Ethnicity Blindness"? - exploring the extremes of recognition performance across categories of faces.

Jeremy J. Tree[1], Alex L. Jones[1]

[1]School of Psychology
University of Swansea
Swansea, UK §

Address for correspondence: J. Tree, School of Psychology, Faculty of Medicine, Health and Life Sciences, University of Swansea, Swansea SA2 8PP, United Kingdom.
E-mail: j.tree@swansea.ac.uk

Declarations Of Interest: None

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement: All data and code used to reproduce the analyses in this manuscript are available on the Open Science Framework: osf.io/yfgvb

Abstract:

The other ethnicity effect (OEE) refers to the common finding that individuals generally perform better in recognizing faces from their own ethnicity than from others. Wan et al., (2017) identified a subset of individuals with a marked difficulty in recognizing other-ethnicity faces, termed *other-ethnicity blindness* (OEB). This study further examines the prevalence of OEB in two large samples of Asian and Caucasian participants, using three analytical approaches to assess face recognition across different ethnic face categories. The first method, based on Wan's percentile-rank approach, additionally adjusted for regression to the mean (RTM), found a 1.9% OEB prevalence, lower than their earlier estimates (8.1% [7.5, 10.6]). Moreover, those identified often displayed *generally poor* face recognition skills. The second approach, akin to a single-case 'dissociation' method (Crawford, 2003), classified just one individual (0.25%) as OEB. The third method defined OEB purely as an *exaggeratedly* large OEE, without using traditional 'cutoff' scores, observed 1.33% of participants exhibited this profile. Bayesian simulations supported these OEB prevalence rates. Overall, the findings highlight the critical importance of accounting for factors like own-ethnicity performance, measurement error and RTM. We also advocate for more conservative classification methods in future OEB research and emphasize that while OEB is rare, it *can* be observed in some individuals. Specifically, adopting the classification of OEB as a 'hyper'-OEE profile may provide a valuable avenue for future research exploration both with respect to those interested in individual variability in OEE and more generally variability in within class recognition performance.

Introduction

The *other ethnicity effect* (OEE) in face recognition research refers to the consistently reported pattern that samples of participants tend to do better at recognising unfamiliar faces from their *own* ethnicity than similar faces from a different ('other') ethnicity (e.g., Malpass & Kravitz, 1969; McKone et al., 2012). However, this difference is observed *on average*, that is, at the level of a group or population, and it is clear that individuals in a population can vary substantially in the magnitude of this effect. Researchers have provided different accounts for this variability including the *contact hypothesis*, which argues that as an individual has an increased level of social contact with a different ethnicity, so their putative OEE for that ethnicity is expected to reduce (e.g., Goldstein & Chance, 1985, Zhou et al., 2019, Ng & Lindsay, 1994). It is also 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). With this OEE individual variability in mind, Wan et al., (2017) examined whether there are individuals who perform so poorly at recognition memory with faces of another ethnicity that they would effectively be *prosopagnosic* for such faces, a pattern they called '*other ethnicity blindness*' (OEB).

This research builds on the topic of individual differences in face processing, which typically focuses on the population 'extremes' of face recognition performance. Assuming performance on a given cognitive task is normally distributed in the population, there are by definition individuals at the tail-ends of this distribution who do very poorly, or very well. Such individuals have been identified and referred to as *developmental prosopagnosics (*DP - very poor - Bate et al, 2019a,b; Bennetts et al., 2022) and *'super' recognisers* (SR - very good - Bobak, Hancock & Bate, 2016; Davis et al., 2016). Wan and colleagues (2017) followed a similar tradition, examining individuals who lie at the bottom end of a distribution of performance for people observing faces of a *different* ethnicity. In all these examples, the *implications* extend

beyond simply being interested in verifying the simple existence of these individuals, but onto issues such as the degree to which such a presentation is *category specific* (i.e., are such individuals also impaired at other visually complex recognition tasks, such as objects) and (with reference to functional cognitive models of face processing) how such poor performance may occur, which in this case must inform our theories of face processing.

As we have established, Wan and colleagues wanted to identify individuals who performed in a manner consistent with 'other ethnicity blindness' (OEB) - that is individuals performing abnormally poorly on recognition memory of unfamiliar OE faces, despite no such difficulties with faces of their own ethnicity. This definition should also reflect an *abnormal difference* between own and other ethnicity performance (or 'exaggerated' OEE) - since, if someone is very poor at recognising *own* ethnicity faces (i.e., *developmental prosopagnosia* - Jansari et al., 2015*)* it would be unsurprising they were also poor with all other types of faces too. Clearly, the operationalisation of the classification of OEB is important. Wan et al., (2017) set out their classification procedure algorithmically: (a) First, they determined a cut off on the measure of other ethnicity face recognition commensurate with performance for the lowest 2% (as determined by a percentile rank) seen in a reference population of individuals *of that ethnicity.* That is, if a Caucasian participant's other ethnicity recognition memory performance was commensurate with the bottom of the distribution of *Asian* participants' ability, they met this first criteria, (b) Secondly, they *excluded* from their key identified OEB sample, individuals who performed in the "*lower-end-of-the-normal-range on <u>own-race</u> recognition ability"* (i.e., below 2% on such faces)*.* This was to remove individuals who perform poorly across the board with faces (i.e., developmental prosopagnosia), (c) Finally they also calculated confidence intervals for the prevalence rate detected with a simulation-based approach, yielding a 'headline' prevalence rate of OEB of 8.1% [7.5, 10.6].

However, using 'cut-off' criteria as an inferential approach has many issues. Namely, these criteria effectively imply a 'fixed boundary' and raises questions around what one does

with individuals who lie very close to these criteria, either just above or below. It is problematic to assume that someone *just* above a cut off is 'normal' (e.g., with own-ethnicity faces) and someone *just* below is 'abnormal' (e.g., with other ethnicity faces). Moreover, if one assumes that most participants will naturally perform comparatively poorly on faces of another ethnicity, if an individual is already poor with own-ethnicity faces, then it is more than likely their OE face performance will fall below cut off. Wan and colleagues acknowledged this and write " *47% of the (OEB) cases [12/22 OEB Caucasian cases and 3/10 Asian cases – see Figure 1 below] have own-race scores in the bottom 15% of own-race abilities ….For these individuals, a small- to moderate-sized OEE (e.g., of only the size of the mean OEE) superimposed on their own-race ability would push them into the clinically impaired range for other-race faces."* That is, of the OEB individuals already identified, approximately *half* performed so poorly with own-ethnicity faces that naturally lower performance for other ethnicity faces could account for their performance classification. Figure 1, adapted from Wan et al. (2017), illustrates these cases. To clarify the potentially confusing switch in the darker and lighter dotted 'cutoff' lines across samples: the cutoffs are determined differently in each instance. The dark dotted lines represent own-ethnicity cutoffs derived from the *same* ethnicity population (e.g., for Asians on the right, this is the cutoff for Asian participants looking at Asian faces). The light dotted lines represent other-ethnicity cutoffs based on populations of the *other* ethnicity (e.g., for Caucasians on the left, this is the same Asian population cutoff) – Asians thus performed better on average with their own ethnicity test compared to Caucasians. In any case, we present Figure 1 to reiterate the point that the OEB cases were generally poor performers. Such that if one classifies OEB as implying someone who is very poor with OE faces but *at least average or better* with own-ethnicity faces, the number diminishes to virtually nil (see the 'dotted' ovals to right of each population bar-chart). In sum, we argue that the original 'headline' figure of 8.1% may be somewhat inflated depending on the degree to which one may wish to take into account how an individual may be performing with faces *more generally* - and thus the *relative difference* of

performance across different face types (or effective OE magnitude), is worth taking into consideration.

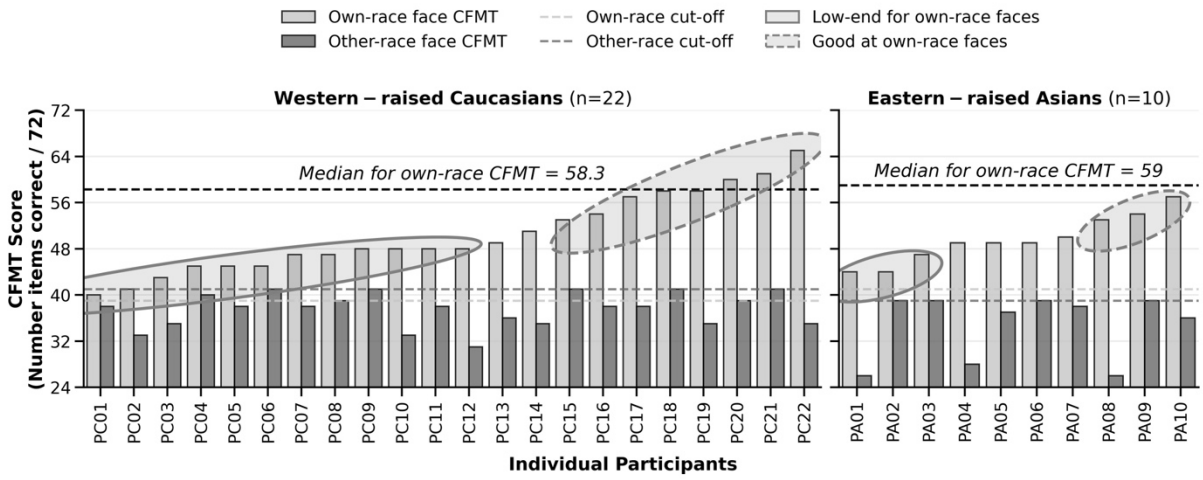

Figure 1 - Wan et al., (2017) - Breakdown of individual OEB cases performance – Note: darker dotted lines represent own-ethnicity cutoffs derived from the same ethnicity population and lighter dotted lines represent other-ethnicity cutoffs based on populations of the other ethnicity.

There is a second issue with the use of cut-off criteria which relates to the statistical properties of measurement. For any given task tested individuals will vary across sampling times, and this variation can reflect measurement error. The classic consequence of this measurement error is *regression to the mean* (RTM): that is, individual performance can vary around its 'true mean', such that an extreme high or low score will naturally move toward the average on its second measurement (Campbell & Kenny, 1999). Moreover, most tests have natural 'floor' or 'ceiling' scoring levels, such that any extreme movement will always likely be in a particular direction. For this reason, if a key group of interest (such as OEBs) is initially selected via 'extremely' poor (near floor) scores on one measure (other-ethnicity face recognition), these same participants might perform better on a *second* measure (own-ethnicity face recognition) simply because of RTM. As a consequence, having *additional* observed

measures is much more preferable. With this in mind, Childs et al., (2021) sought to determine whether individuals who perform particularly poorly/well on faces of their *own ethnicity* showed OEE of a different magnitude, and the authors took careful steps to recognise the issue of RTM by using three different face recognition measure observations. They reported that the OEE remained constant across all levels of own ethnicity performance, consistent with other word that indicates that performance with own ethnicity faces tends to be the strongest predictor of performance with other ethnicity faces (Cheung, Quimpo, & Smoley, 2024, Chen, Kassa, & Cheung, 2023, Trawiński, Aslanian, & Cheung, 2021). As a consequence, it seems unlikely that individuals who are OEB are necessarily likely to reflect a specific level of *own-ethnicity* ability. Moreover, this finding is observing performance at the aggregate level of a *population mean*, and thus it remains unclear around the prevalence of singular individuals that may fit criteria for OEB. Unfortunately, in the key study by Wan et al., (2017) no third test option was available, and thus what the observed prevalence of OEB might be with such an approach is unknown. Here, we seek to explore the prevalence rates of OEB with a large data set (N=400 Caucasians and N=424 Asians) who completed three face recognition tests (taken from Childs et al., 2021). This enabled us to identify key candidates in a manner that we can more systematically deal with the issues around RTM. We also sought to determine what proportion of this identified sample scores above average on own ethnicity faces (to rule out generally poor face recognition memory - see above). As a consequence, using a much larger sample data set we sought to confirm if we could identify prevalence rates of OEB equivalent to that of Wan et al., (2017) using a largely similar approach with some additional methodological benefits.

Another issue with the cut off approach goes beyond issues of RTM and relates to the degree to which one chooses to determine the criteria for ‘normal’ own ethnicity face performance. As we have stated, it is disingenuous to interpret performance just above cut off as normal, given so many identified OEB cases were largely generally poor with faces. Moreover, the approach is focused on interpreting one performance pattern (other ethnicity) as

'impaired' and another (own ethnicity) 'intact', where the latter essentially rests on interpreting a null effect (which is notably problematic). To deal with this issue, one option is not just to consider scores on each given test *independently,* but in addition to consider the *relative difference* between tests within individuals (Crawford, Howell & Garthwaite, 1998). Under this *single case* approach, an OEB individual must: (a) have another ethnicity mean score consistent with 'impairment' relative to their reference population (using a modified *t*-test - Crawford, Howell & Garthwaite, 1998), (b) have an own ethnicity mean score consistent with 'intact' performance (using a modified *t*-test) *and* (c) display a difference between these two measures (i.e, OEE) that is much larger than that typically observed by the population (i.e., an *exaggerated* OEE, or within category 'dissociation') using a test of difference (ZDCC) from the Revised Standardized Difference Test (RSDT). Under the definition of Crawford (2003), such an individual would meet the criteria for a putative 'classical dissociation', and this approach has been used in the identification of individuals who meet the criteria for *developmental prosopagnosia* (i.e., very poor face recognition ability; see Fry, et al., (2022)). Additionally, this third criterion incorporates a vital but often ignored parameter - the *correlation* between the two observed tasks. That is, individual performance across a face recognition measure (regardless of type of faces) will of course be related, and taking this into account is essential when interpreting individual cross-category differences. Thus, this secondary approach incorporates *all* elements of an individual's performance across face types, rather than looking for singularly poor performance with faces of the other ethnicity. In sum, one of the approaches of this paper is to point to the fact that a critical individual differences goal in this context is not to simply determine the 'cut-offs' for a distribution of task performance in a reference population (i.e., to determine 'extremes' of task X or task Y), but rather to determine the 'extremes' of *relative* performance across two tasks (X vs Y) that are of interest in the context of interpreting ' dissociations'. In addition, we will also consider identifying OEB as simply an 'exaggerated' or hyper-OEE effect - in fact, McIntosh (2018) argues that one can classify individuals with key

‘dissociations’ *entirely* on the basis of the magnitude of the between-task difference. In this context then, one could ignore the ‘cut-offs’ classification for each task (to side-step its related problems) and ask - how many individuals appear to present with a ‘hyper-OEE’, as defined by a within-individual relative difference between own/other faces that is beyond that typically observed in the population? (see Methods – Analytic Strategy for more in depth details of these approaches).

A final set of issues that underpins these single-case classifications is the handling of uncertainty in estimation of the parameters needed to conduct the tests, and that there is no formal statistical model that describes how the observed data was generated. In the first case, consider that the statistics used in the RSDT are point estimates (the sample mean or correlation, Crawford, 2003), and the extremity of a single score is calculated relative to those values - thus, uncertainty in those values impacts whether a single score may or may not be classified as extreme. Bayesian inference naturally handles uncertainty in estimates (Kruschke & Liddell, 2018), and thus provides a useful vehicle for characterising uncertainty in these estimates in the form of the posterior distribution, which fully represents the plausible values these statistics might take given the current data, and a model of that data. The second case is directly related to the first - both the percentile approach and the RSDT make no explicit claims about how the sample data was generated. That is, what process might give rise to the observed data? A useful statistical model for this sort of data is the multivariate-normal distribution, which is parameterised by means, standard deviations, and correlations amongst measures, exactly the estimates needed to calculate tests like the RSDT. By applying this model to data and using Bayesian inference to quantify the uncertainty in the estimates of this model, we are able to generate new datasets that have similar properties to the observed data but propagate the uncertainty through them. This *posterior predictive* simulation (Gelman, Men, & Stern, 1996) allows us to infer the likely rates of OEB in large sample populations by generating many new datasets and conducting our analytic strategy many times over.

All these approaches will enable us to determine the prevalence of OEB, to the extent that we will confirm if such individuals are even observed in principle – and our work will provide useful guidance for their future identification. But there also follows what such observed behaviour *implies* for our cognitive models of face processing. We have already mentioned that there is a quite established field of research around 'extreme' individuals who have 'extreme' good/bad *general* face recognition performance (DP/SR), but the presence of individuals with OEB would indicate that *within* a category of complex stimulus type there may well be similarly 'extreme' divergence ('dissociations') of individual level performance. It is fair to say that just as the work on DP/SR has informed our cognitive models of face processing, little attention has been made to the possibility that information processing *within* a stimulus category may equally extend across a considerable individual differences range. Any models or theories of visual learning need to account for such evidence if a presentation like OEB is established – this could be via testing of individuals with OEB with other instances of measuring *within-class* discrimination ability (such as other objects) in order to determine whether performance is category specific, and speak to a long running debate around the 'special' nature of faces as a class of stimuli – (see Towler & Tree, 2018).

In addition, confirming the observation of OEB candidates has significant implications in the 'real world' beyond the modest *average* OE typically observed in studies. Wan and colleagues point out that in legal contexts, assessing eyewitness accuracy for other-race suspects should involve evaluating the specific witness's face recognition skills. The presence of OE individuals would thus suggest that although the average OE effect may be minor, some individuals are severely impaired, which can impact both eyewitness testimony and social interactions. This variability, which we seek to re-affirm, implies that some people might be extremely unreliable as eyewitnesses in other-race identification, even under optimal conditions, with this practical implication being currently ignored particularly since it could be 'masked' by typically average (or even above average) *own* ethnicity performance for that same individual.

Moreover, previous work with DP individuals suggests these challenges can have significant negative impacts on everyday social interactions (Burns et al., 2023), and thus the analogy with OEB individuals would likely be similar, but for a set of social interactions between *specific* colleagues say in the workplace. Repeatedly failing to recognize others in this case could lead to awkward situations and misunderstandings, with those not recognized potentially misinterpreting the behaviour as racially motivated rather than as a perceptual issue. Thus, the confirmation of OEB would thus warrant future research that systematically investigates the range of individual impact of other-race face blindness on daily social and workplace interactions.

In summary, the current work explores the prevalence rates of individuals who present with extremely poor unfamiliar face recognition with other ethnicity faces as compared to own-ethnicity - a presentation dubbed 'other ethnicity blindness' (OEB), which existing literature suggests may be as high as 8.1% (Wan et al., 2017). Here, we seek to do several things: firstly, with a new and larger sample, we aimed to identify and report proportions of OE cases using the methods of Wan and colleagues whilst also using an approach that attempts to alleviate some of the likely consequences of measurement error and RTM discussed earlier. Secondly, we aimed to determine proportions of OEB using a novel approach which would define such a profile as a 'classical dissociation' across face types (Crawford 2003). Our approach in this latter case points to the fact that a critical individual differences goal in this context isn't simply to determine the distribution for singular task performance in a reference population (i.e., to determine 'extremes' of task X or task Y), but rather to also determine the 'extremes' of *relative* performance across two tasks (X vs Y) that are of likely interest in the context of interpreting 'dissociations'. To foreshadow our subsequent findings, we demonstrate that this more comprehensive (and arguably conservative) second approach indicates that the likely proportion of such OEB cases identified is very small, *but not zero*. Finally, given this small proportion identified, we adopt a third analytical approach that focuses entirely on the *relative difference*

between own and other ethnicity scores (inspired by McIntosh, 2018); to understand the proportion of individuals who indicate a profile consistent with a 'hyper' OEE, and thus further explore the extremes of the OE effect itself. Finally, we combine these approaches with a Bayesian statistical model and use posterior predictive simulation to infer the likely rates of OEB cases in large-sample populations.

Methods

Participants

Sample	***Country***	***Age Mean***	***Age SD***	***Woman, Man***	***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

Table 1 - *Participant count, age means and standard deviations for our samples*

All data considered in this work is taken from an earlier study by Childs et al., (2021). These data comprise two large samples of participants - N=400 Caucasians (from three countries, Serbia, the UK, and Australia) and N=424 Asian individuals from Japan, Korea, China, and Singapore. - Notably, these samples are twice the overall size of those reported in Wan et al. (2019). Sample key demographic information is presented in Table 1 above.

Materials:

		Australian		*Boston*		*Asian*	
Country	N	Mean	SD	Mean	SD	Mean	SD
Australia	102	55.15	7.47	55.94	7.87	50.9	7.88
Britain	195	54.37	7.66	55.19	8.45	52.2	8.61
Serbia	103	57.69	7.35	58.14	8.61	51.04	8.22
China	103	48.73	7.52	47.32	8.9	56.59	8.31
Japan	120	48.78	7.41	51.91	7.49	56.74	7.5
South Korea	109	52.73	8.3	51.5	8	55.5	8.98
Singapore	93	50.65	8	49.08	8	55.11	7.55
Caucasian	400	55.42	7.64	56.14	8.41	51.57	8.33
Asian	424	50.19	7.95	50.08	8.28	56.03	8.11
Total N	824						

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 2 – *Descriptive statistics for each country cohort on CFMT measures.*

Since we sought to explore the 'extremes' of the OEE effect in unfamiliar face recognition, it was critical we used a well-validated measure of face recognition ability – the ideal candidate is the Cambridge Face Memory Test (CFMT). In this case, we have three established versions: *Boston* (Duchaine & Nakayama, 2006), *Australian* (McKone et al., 2011), and *Asian* (McKone et al., 2012). Previous work has observed all three of these tests have high internal reliabilities (Horry et al., 2015; McKone et al., 2012), and our own work observed similarly high Cronbach alphas: Boston CFMT a = .917, Australia CFMT a = .873; and Asia CFMT a = .846. These tests and the procedures are described in detail in the work of Childs et al., (2019). But a short summary is provided here – (a) each of these computer tasks were presented to our participants for online testing using a bespoke programme constructed by the department's software technician, (b) the order of CFMTs was counterbalanced for each participant to reduce order effects and (c) following completion, participants were thanked for their time and awarded course credits. Each of the CFMT tests have been described extensively elsewhere (see references above), but in brief the experimental procedure follows the original design by Duchaine and Nakayama (2006). The task involves three phases using greyscale images of men's faces with the hair removed.

1. **Learn Phase (18 trials)**: Participants are shown three target faces from different angles (left, front, right) and must identify the target among two distractors.
2. **Novel Phase (30 trials)**: Participants identify target faces under different lighting or viewpoints, again in a triad with two distractors.
3. **Noise Phase (24 trials)**: This is similar to the Novel phase but with Gaussian noise added to increase difficulty.

Between phases, the six target faces are displayed for 20 seconds as a reminder. The participants' accuracy across all phases is recorded and summed to assess recognition ability.

Each test was presented in three different orders, but no significant differences were found between the orders (see Childs et al., 2019).

Importantly for our purposes, all three CFMTs have been used extensively in face testing and studies of the OEE with robust effects sizes reported (e.g., Zhao et al., 2014; Zhou et al. 2019), including our own work. Descriptive summary scores are shown in Table 2. In the work of Wan et al., (2019) they used two of these CFMT tests, the *Australian* and the *Asian.* However, given we have test scores on the CFMT for all three formats, this enables us to identify extreme OEE performance via two *different comparisons* (unlike the work of Wan et al., 2019), to deal with issues around measurement error and RTM (discussed earlier, see analysis strategy below). Put simply, if a case of OEB were identified via a particular own/other CFMT paired comparison (e.g., Asia vs Australian), does this pattern remain when observed by *another* CFMT paired comparison (e.g, Asia vs Boston)?

Analytic Strategy:

We sought to explore the prevalence of cases of OEB using three different analytical approaches. Firstly, we followed a similar approach that draws on the key work of Wan and colleagues and subsequently explore the degree to which these observations survive multiple observations. Secondly, we used an approach inspired by single case analysis proposed by Crawford, Howell, and Garthwaite, (2005). Third, we used an approach that specifically focuses on identifying cases solely on the basis of their *disproportionately large* OEE, inspired by the proposals of McIntosh, (2018). Finally, we extend these findings using a formal statistical model, the multivariate normal distribution. Using Bayesian inference, we conduct a posterior predictive simulation study demonstrating how this approach captures the likely rate of cases in large-sample populations. We outline each approach in turn below. To foreshadow what is to come, we demonstrate that *any* approach that attempts to identify prospective OEB candidates (or

other forms of ‘extreme’ performers), would do well to undertake multiple testing observations, in order to have more confidence in the stability of ‘impaired’ (or otherwise) performance. With this in mind, in our final section, we return to the issues of using ‘cut-offs’ and measurement error, by discussing the dangers of ‘spurious’ dissociations, which again illustrates the challenge of drawing inferences from few observations.

Wan et al. (2017) approach –

As a first step, we followed the procedure for classifying OEB cases suggested by Wan and colleagues as closely as possible. The authors used a ‘cut-off’ approach, based on percentile ranks, to determine individuals who performed at a level consistent with the poorest 2% of the population on the measure of other ethnicity using the CFMT Australia (for the Asian sample) and the CFMT Asia (for the Caucasian sample). The authors point out that although many studies typically use a ‘cut-off’ based on Z-scores (e.g., a score that is equivalent to the reference sample mean, minus 2 standard deviations), this is inappropriate when the measure is not normally distributed (Crawford, Garthwaite, & Slick, 2009). Given accuracy scores for the CFMT are typically non-normal (Degutis et al., 2023) and often negatively skewed (which they observed in their own data), they adopted the alternative percentile ranks approach, calculated from every score in the sample. Using these ‘cut-offs’, Wan and colleagues identified candidate OEB cases *excluding* any individuals who performed below cut-off on measures of their own-ethnicity (i.e., were generally poor with faces regardless of ethnicity). In short, we undertake a similar approach for both our Caucasian and Asian samples, using the same comparators (CFMT Australian and Asia), to provide equivalent prevalence rates to their work.

However, given the consequences of measurement error (which was estimated to be as large as 5-6 items out of 72 on the CFMT by Wan and colleagues), and the impact that would

have on individuals falling above or below these 'cut-offs'. We sought to adopt a second approach that illustrates the impact of likely measurement error and RTM, given we *also* have performance on the CFMT Boston. In this case, after identifying candidates from our OEB Asian sample using their OE pattern across own (Asian) and other (Australian) testing (following Wan and colleagues), we *repeated* the same approach across another own (Asian) and other (Boston) testing pairing. That is, we were able to undertake this comparative process *twice* – and we sought to *only* classify individuals who met the above criteria across both comparative analyses. Finally, having undertaken these analyses and reported the number of individuals who meet criteria, we also reflect on these candidates' *own-ethnicity* face performance. That is, we consider how many of this sample have performance on own-ethnicity faces that would be commensurate with average or above performance. As was mentioned in the introduction, many OEB cases identified by Wan and colleagues generally had quite poor own-ethnicity face performance (see Figure 1) and we seek to determine if we observe a similar pattern – to determine if all cases we identified may simply be explained as generally poor performers.

Crawford (2003) Analysis Approach

As an alternative means of identifying individual cases of OEB, we used an approach that aimed to find a 'classical dissociation' of performance between own and other ethnicity performance (e.g., Caucasian CFMT and Asian CFMT). As discussed earlier, a range of statistical tests have been developed to compare single cases against matched control samples, to estimate how unlikely it would be to find performance more extreme than that of the single case, relative to the normal population distribution. Crawford, Howell, and Garthwaite (1998) adapted the paired t-test to create a parametric approach for comparing the differences between a patient's performances on two tasks relative to the distribution of paired differences in control participants. They further developed the Revised Standardised Difference Test (RSDT), which standardises scores on each task before assessing the differences (Crawford & Garthwaite,

2005). We follow this approach by applying the criteria suggested by Crawford et al. (2003) and followed the implementation of these tests developed in their software programs, Singlims_ES.exe and RSDT_ES.exe (Crawford et al., 2010). To meet criteria for a 'classical dissociation' an individual case had: (a) performance on the other ethnicity face test that differed significantly from that of the normal population, which was estimated based on a reference sample of individuals who match that ethnicity – this was via the Crawford t-test approach, and (b) the *difference* in performance of that person across own/other ethnicity face tests must differ significantly from the difference scores of the normal population on the same two tasks – in this case then the reference samples all matched the ethnicity of the participant. Using these reference samples we calculate the test population mean/standard deviation and the correlations between paired tests which are used in the Revised Standardized Difference Test (Crawford & Garthwaite, 2005), which calculated an effect of OEE that would reflect a difference (Z_{DCC}) of a magnitude that was statistically significant from that expected within individuals for the population comparison in question (e.g., Boston versus Asian or Australian versus Asian). Finally, in order to recognise the likely potential issues of measurement error and RTM, this process was undertaken *twice* across two own/other ethnicity comparisons – i.e., Boston-Asian *and* Australian-Asian – and a key candidate would have to fit criteria *in both cases*.

McIntosh (2018) Analysis Approach

As a final approach, and to both side-step issues of using classifications specifically linked to 'cut-offs' as well as move away from an approach that would likely draw on low own- ethnicity face performance, we sought to identify individuals who showed a 'Hyper-OEE'. This was inspired by the proposals of McIntosh (2018) who argued that a simpler approach for identifying 'dissociations' is merely to identify individuals who perform in a manner that is *disproportionately* poor on category X versus Y. This was following the conclusions of Crawford & Garthwaite, (2007, p362) whose work identified the abnormal large difference as the *"most important*"

component when they characterised the approach we discussed earlier (see Crawford (2003) analysis approach above).

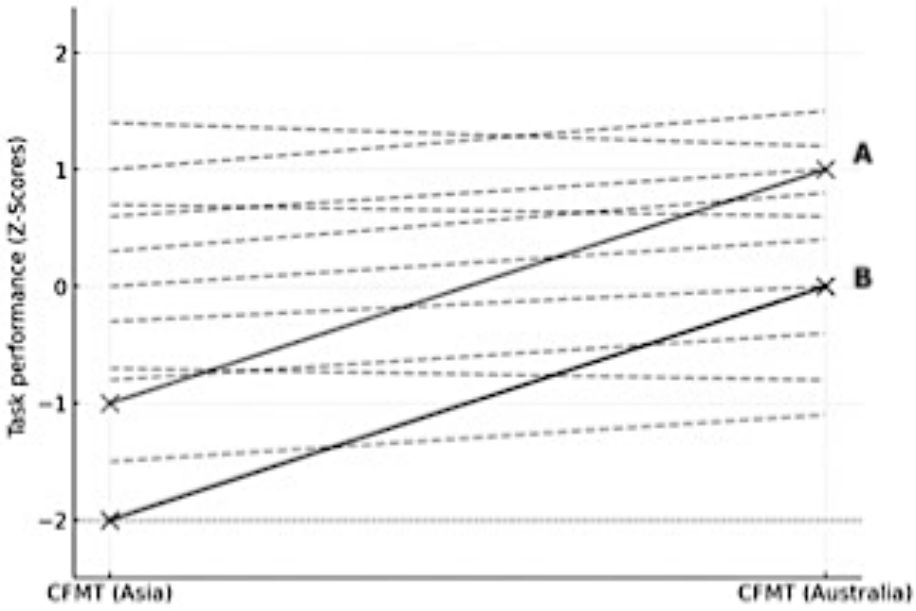

Figure 2 – hypothetical example of two (A/B) potential OEB cases following McIntosh (2018) proposals – the dashed lines represent other observed performance of other participants.

McIntosh's point is that if the purpose of identifying dissociation candidates is to observe individuals who show abnormally *large differences* across tasks (captured by a Z_{DCC}), then why might you solely focus on individuals who *also* show severe impairment on X, particularly when that can often entail naturally lower performance on Y? For example, in Figure 2 above we illustrate two hypothetical OEB Caucasian cases (marked A and B) both of whom have a Z_{DCC} of 2, along with a number of illustrative 'controls' (marked by dotted lines) – in one example (B) this individual would also meet the Crawford criteria discussed above, with additionally 'impaired' performance on the other ethnicity test (CFMT Asia) of -2SD performance. In the second hypothetical cases (A), the magnitude of Z_{DCC} is identical, but because the individual is a *generally* better performer they wouldn't meet criteria under the Crawford approach discussed

earlier. McIntosh (2018) undertook a number of empirical simulations which showed that focusing only on candidates who present with a significantly extreme difference between tasks, increases power to detect dissociations, with transparent and stable control of Type I error – critically, the correlation between the two observed scores plays an important part, since naturally a large divergence between observed scores becomes increasingly unlikely when scores are increasingly highly associated (regardless of the general level of ability). Moreover, the jettisoning of the additional criterion that X must meet a 'cutoff' of -2SDs also eliminates the possibility that individuals very near to this 'line in the sand' are somewhat arbitrarily ignored. Since the objective of our work is to potentially identify individuals who are OEB, it would be interesting to characterise such a presentation as a 'Hyper-OEE' (i.e., solely a large $Z_{DCC)}$, so that candidates such as the two outlined above in Figure 2 would meet such criteria.

In this case then, the focus is only on using the RSDT approach, to calculate individuals who show an extremely large OEE (i.e., statistically significant Z_{DCC}). Since the OEE is a phenomenon we expect to see across any population of individuals tested regardless of ethnicity, the samples were combined entirely to calculate the relevant means, standard deviations and correlations, allowing us to examine whether any given participant exhibits an abnormally large difference. Moreover, as in previous instances, when identifying candidate cases of 'Hyper-OEE' performance, individuals would have to show disproportionate differences for *both* own/other ethnicity test comparisons, that is we take the McIntosh approach even further by demanding that the unlikely observation occurs *twice.*

Bayesian posterior predictive simulation

While the single-case approaches described above are useful, especially when observed over multiple tests, there are some drawbacks. The first is that they rely on point estimates calculated from sample data (e.g., means, correlation coefficients), which ignores inferential uncertainty in these estimates. This is a key issue here, as outside of measurement error and RTM,

uncertainty in sample estimates also impacts the calculation of individual Z-scores, difference scores, and p-values that allow us to classify individuals as impaired or otherwise. For example, a slight difference in the mean impacts the Z-score of a single score, which further impacts its single-case p-value, and so on. Second, the analyses make no clear assumptions about how the actual sample data was generated - that is, there is no formal statistical model involved in these single-case tests, which makes them relatively descriptive as opposed to inferential. The third issue is that the tests are limited solely to the sample we collected, and thus our observed rates are proportions of our sample. Given these data, what might be the typical rate observed in a larger population? We apply Bayesian inference to address these limitations, through the use OF statistical modelling. We use a multivariate normal distribution to fit to the sample data (Caucasian and Asian, separately), which simultaneously estimates the means, standard deviations, and correlations amongst measures, which are the key statistics used in the calculation of the single-case p-values. Importantly, through the application of Bayesian inference, we obtain inferential uncertainty in these estimates via the posterior distribution, which describes the plausible values these estimates may take, given the data and the applied model. We take this one step further by drawing a posterior predictive simulation from that model - this distribution feeds the uncertainty into the estimates into a model and generates many large-sample datasets that resemble the collected data. In these large sample datasets, we repeat our three approaches that allow us to estimate a likely distribution of OEB prevalence in the population.

Transparency and Openness Statement

The full codebase and data used in this manuscript are available at the Open Science Framework: osf.io/yfgvb. The analyses in this work were not preregistered. The materials for this work were taken from Childs et al (2021), the materials for which are available from this OSF repository: osf.io/bwhtg.

Results

We present our prevalence findings separately for the three different approaches described earlier.

Wan et al. (2017) approach -

As a first step, we calculated percentile 'cut-offs' scores for each of our three CFMT tests, which were similar to existing work - Boston = 37/72 (rank 1.4%), Asia = 38/72 (rank 1.7%), Australian = 39/72 (rank 1.6%). These 'cut-off' scores were generated with reference to samples of own ethnicity populations (i.e., Boston/Australian cut-off using the combined Caucasian samples and Asia cut-off using the combined Asian samples). Under the approach of Wan and colleagues (using just the Asian/Australian CFMTs), OEB candidates were expected to perform below this cut off for the test of other ethnicity faces, whilst performing above cut off for the test of own ethnicity. With this in mind they found 5.7% of their Asian sample met these criteria using the CFMT-Australian and 9.7% of their Caucasian sample met this criteria for the CFMT-Asia – making a grand total of 8.1% of their sample OEB (the 'headline' in their abstract). Following this same process 8.5% (95%CI [5.84%, 11.14%] - N=36/424 – see Appendix 1) of Asian participants and 4.75% (95%CI [2.67%, 6.83%] – N=19/400 – see Appendix 1) of Caucasian participants met these same criteria – with a grand total of 6.67% (55/824) of our sample being OEB. Overall, it is apparent that using the same approach we find a somewhat similar 'headline' figure, albeit a bit lower and with the total percentages of Asian/Caucasian participants meeting 'criteria' being flipped in comparative magnitudes. This likely reflects the fact that in their work although the 'cut-off' for the CFMT-Australia matched ours, in their case the CFMT-Asia 'cut-off' was higher (41).

Participant ID	*Participant Country*	*Age*	*Gender*	*Boston Total*	*Aus Total*	*Asia Total*	*Asia Percentile Rank*
35wj7E6m	China	20	Man	33	31	39	2.40%
h42jFq83	China	19	Man	36	38	40	3.70%
40E8r5HS	S Korea	20	Man	33	33	42	6.00%
1yEn750C	Japan	19	Man	32	36	46	12%
9MZd712y	China	18	Man	37	34	46	12%
B112Kv3u	China	20	Man	35	36	47	15%
52FAV6d9	Singapore	21	Man	30	33	49	21%
5iv74HR7	Japan	20	Man	37	39	57	52%

Table 3 - breakdown of 9 Asian OEB individual cases – Note: in all cases individuals meet criteria for 'cut-off' impairment on both other ethnicity tests.

However, in addition to the problem of 'cut-off's often moving depending on the observed samples, a key problem we have argued is that for one to truly interpret scores as reflecting prospective OEB 'candidates' it is important to attempt to rise to the challenge of measurement error and the impact of regression to the mean (RTM). Given we have two 'other-ethnicity' tests for our Asian (N=424) sample (Australian/Boston), we are able to use both tests to determine the number of individuals who are impaired at *both*. For full transparency, in Appendix 2 we present all Asian cases who performed below 'cut-off' on either only the Australian CFMT, or only the Boston CFMT or both tests – and this provides details of the overlap between these two OEB identified groups. That is, we could determine what proportion of individuals who meet OEB cut off criteria on *both* – under this approach the observed prevalence is much lower, with the rate of OEB in the Asian sample being 1.9% (N = 8/424).

Table 3 above illustrates further features of the identified Asian subsample. First, all cases were men, and second, many cases performed poorly on the *own-ethnicity* CFMT-Asia,

with all but one performing in the lower quartile on this test. We highlight this because Wan and colleagues asked the question, might all OEB cases identified via their approach simply reflect generally low face processing ability; and thus if one also assumes a modest OEE is present this would tip such individuals over the edge into meeting the 2SD 'cut-off' on other ethnicity faces. In fact, in their consideration of this possibility, they focused on individuals who performed "*in the top 50% of own-race abilities (i.e., better than the median)"*, and observed 3 candidates met such a performance level (see Figure 1) or 3/444 (0.67% of the sample). In our case we find a similar situation, with only *one* (1/424 - 0.24%) individual in our Asian sample (case 5iv74HR7) meeting such criteria - see Table 3 above marked in bold. Overall, across our work and theirs, although generally poor performance cannot explain *all* cases identified using the approach of Wan and colleagues, such an explanation seems to capture performance of the near majority. In sum, following their approach, our analyses indicate a lower rate of OEB than the "8.1%" headline of Wan and colleagues, but we attribute this to our stricter criteria, since it is clear that any observed rate clearly falls dramatically if confirmation is demanded consistently (i.e., more than once).

In sum, we would urge some caution in accepting the initial figures for rates of OE 'blindness' (e.g., McKone's headline 8.1% proportion), and on the basis of our first analysis findings if one follows the spirit of their approach and considers RTM, a more plausible rate of ~2% is observed. At first blush this appears to mirror reported rates of developmental prosopagnosia reported in the literature (i.e., individuals with very poor *own-ethnicity* ability (Kennerknecht et al., 2006), although in that case the 2% in their case may simply reflect the use of a 2SD 'cut-off' on *a single test*. Despite this, it is also apparent that most in this 2% OEB sample (see Table 3), are not particularly good with *own-ethnicity* faces, and thus one may wish to caveat this 'OE blindness' observation by stressing that the majority of presented cases may simply be explained thus. However, it is fair to say that an approach that demands an OEB

candidate must have average or above own-ethnicity performance may be overly conservative, since this would entail an overall other-ethnicity effect (OEE) of a magnitude that is perhaps unrealistic. As a consequence, we sought to identify individuals who perform poorly with other-ethnicity faces *and* show a comparative difference *across* their own/other ethnicity ability that was highly unlikely to be seen in the general population (moving beyond a simple cut-off approach). This motivated our utilisation of a second approach to classify OEB.

Crawford analysis

Crawford, Garthwaite & Porter (2010) provide criteria that serves as an excellent means of operationalising the identification of single case 'dissociation' patterns, akin to a neuropsychological *case series*. Here, we report each single individual who meets all elements of these criteria. This analysis utilises the means, standard deviations, and correlations between measures to identify individuals with notable dissociative patterns. In addition, rather than using the 2% ranked percentile cut-off suggested by Wan and colleagues above, we will use a single case deficit Crawford modified t-test for comparison of own/other ethnicity test scores, which will establish a statistical difference between the participant's score and that of the reference sample in each case.

Taken together, our second analysis was as follows: Step 1) identify individuals who perform very poorly with other-ethnicity faces (modified t-test) and normally on their own ethnicity faces (modified t-test). Step 2) of this subset, how many show a difference across own/other ethnicity performance (using RSDT) such that this is so *disproportionate* it would equate to less than 2% of the population (i.e., something akin to a *category specific* impairment in classical neuropsychology). In addition, given we have three CFMT tests we could confirm this key dissociation across *two* different pairwise comparisons.

We focus on the Caucasian sample[1] for this analysis. In *step 1*, we sought to determine if the participant was significantly impaired at the CFMT- Asia, and to maintain consistency with the initial Wan et al analysis above, we used the mean and standard deviation information provided by the Asian *reference sample* (CFMT-Asia = 56.03, SD=8.11). Based on the Crawford *t*-test, a score of 40 or less on the CFMT-Asia would be evidence of a statistically 'impaired' level of performance. This is clearly higher than the percentile rank approach described earlier, where cut off was set as 38. After eliminating all other candidates, in *step 2,* we use the mean and standard deviation accuracy performance on each of the three tests provided by the Caucasian *reference sample* - in addition we calculated the correlation between Boston/Asia = 0.582 and Australian/Asia = 0.524.

With a single pairwise comparison between tests (e.g. Boston to Asia CFMT), several candidates for OEB appeared. However, this pattern was generally not confirmed with a second pairwise comparison. We detected only a single participant (0x065PFT) meeting criteria for a *category specific* impairment in both pairwise comparisons: Boston (58) to Asia (38), RSDT=2.66, p=0.008, Z-DCC=2.674, Australia (58) vs Asia (38), RSDT= 2.616, p=0.009, Z-DCC=2.62. Consequently, characterising OEB as a presentation in which individuals have impaired performance with OE faces that would equate to a dissociation, there are virtually no observations meeting this criteria (1 out of 400, or 0.25%).

Overall, if we use a Crawford et al., (2010) approach to determine whether any of the individuals in our large Caucasian sample present with a *classical dissociation* of severely impaired other ethnicity ability that would be consistent with a category specific problem, we are able to identify only a *single* individual (0x065PFT). This appears consistent with our earlier

[1] In this Caucasian sample investigation, we have two matching *own* ethnicity samples from the Boston/Australian and one different ethnicity sample (i.e., for Asia) - and thus in this case we have reference samples that are matched to the participant ethnicity and one that is not - for the Asian sample the reverse is true - as a consequence we focused solely on Caucasians because in the reverse scenario the *majority* of our reference samples are of a different ethnicity (i.e., 2/3) which may be problematic.

analysis, when we sought to determine candidate OEB individuals who performed at/or above average with own-ethnicity faces (see above), where again only *a single individual* was identified in the Asian sample (5iv74HR7 - see Table 3). In short, if we use these definitions as the basis of OEB it is unlikely we will find many (if any) such individuals, particularly when takes such conservative steps to consider measurement error and RTM. However, having adopted two different approaches to consider the issue of OE blindness, it is apparent that what is crucial is that they should at least show a *disproportionately large* OE effect, whilst also having performance on the OE category that is 'impaired' (below cut-off). Which begs an interesting question, how often do single individuals show such 'hyper-OE' effects, *regardless of whether their other-ethnicity performance is 'impaired' (i.e., below cut-off or statistically different)?* Put simply, what if we choose to identify OEB in a manner that does not require they must also be below a 'cut off' on other ethnicity faces - after all, in the case of both the single individuals we identified in our two earlier analyses (0x065PFT and 5iv74HR7) each had an effective OEE for Boston/Asia and Aussie/Asia of 18-20 points, which is 4-5 *times greater* than that seen against the population average. In the next section we thus use an alternative means of classifying OEB, where in this case OEB is defined as individuals who appear to show an 'abnormally' large OEE or 'hyper' OEE.

McIntosh approach

In the previous section it appears very few individuals meet the criteria of a 'classical dissociation' across own and other ethnicity faces such that they fit the criteria outlined by Crawford et al., (2010). Nonetheless, as we have established there *are* individuals who demonstrate a *disproportionately large* OEE or a 'hyper' OEE. With this in mind, McIntosh (2018) provides the inspiration for a third analytical approach that focuses specifically on identifying individuals who present with a 'hyper' OEE. In this case, using the Revised Standardized Difference Test (RSDT) we calculated an effect of OEE that would reflect a

difference (Z_{DCC}) that was statistically different from that expected within individuals for the population comparison in question (e.g., Boston versus Asian or Australian versus Asian). In this case we combined the entire samples, to provide global means and standard deviations for each of the tests: Boston, $\bar{x}$=53, sd=8.87, Australian, $\bar{x}$=52.73, sd=8.22, Asian, $\bar{x}$=53.86, sd=8.51- and correlations between these tests - Boston-Asian r=0.445, Australian-Asian r=0.413. In the earlier examples we *pre-selected* individuals from our Caucasian/Asian samples on the basis of the fact that their performance on other-ethnicity faces met the criteria of an 'impairment' (as so defined by a percentile rank or modified t-test). Our approach here is different in that we consider only whether a sizable difference between tests (significant ZDcc), accounting for their correlation, might flag a 'dissociation' (McIntosh, 2018), regardless of absolute performance on either task. Similarly to previous examples, we can repeat this comparison twice using the parallel forms of the Boston and Australia CFMT formats.

Participant ID	*Ethnicity*	*Country*	*Age*	*Gender*	*Boston Total*	*Aus Total*	*Asia Total*	*Boston - Asia: ZDCC*	*Boston - Asia: p-value*	*Aus - Asia: ZDCC*	*Aus - Asia: p-value*
6dAX0b00	Caucasian	Australia	19	Man	66	65	47	2.27	0.032	2.3	0.034
608EX4ih	Caucasian	Serbia	19	Woman	58	58	38	2.43	0.022	2.51	0.021
Uy465ty8	Caucasian	Serbia	19	Woman	57	58	40	2.08	0.049	2.27	0.037
6rLA0u75	Caucasian	Serbia	22	Woman	64	64	45	2.28	0.031	2.41	0.026
Zh177V0f	Caucasian	Serbia	19	Man	68	62	44	2.85	0.007	2.29	0.035
0x065PFT	Caucasian	UK	21	Woman	60	57	39	2.54	0.016	2.27	0.037
5EM416iV	Asian	China	18	Woman	43	45	64	-2.32	0.028	-2.13	0.05
1NU16e6y	Asian	China	18	Woman	49	48	69	-2.23	0.035	-2.35	0.03
b9aX196i	Asian	China	18	Man	41	43	68	-3.01	0.004	-2.84	0.009
17rM5t2Q	Asian	Japan	19	Man	41	37	60	-2.07	0.05	-2.63	0.015
19v9qz7Q	Asian	Singapore	18	Woman	45	44	64	-2.09	0.048	-2.25	0.038

Table 4 - Key individual cases showing a 'hyper' OEE – Note: in each case both pairwise test comparisons yielded a Z_{DCC} *sizeable enough to be statistically significant.*

Using this approach, we identified six Caucasian and five Asian participants who show a difference between own/other ethnicity testing that was statistically significant (RSDT) and this pattern was observed *twice,* and these 'hyper-OE' candidates are presented as a case series in Table 4 below. That is, we have a prevalence rate of 6/400 (1.5%) for Caucasians, 5/424 (1.2%) for Asians and 11/824 (1.33%) overall. In Figure 3 (top), we focus on the Asian participants and plot performance for each of the six participants similarly to Wan et al (see Figure 1). The horizontal line in the top figure indicates performance that would constitute a level commensurate with a ranking of 15% or higher in the general population on own-ethnicity (Asian) faces. Overall, not only is it clear that several in this group are doing *extremely well* with own-ethnicity faces (i.e., 'superior'), in nearly all cases individuals are performing in the bottom quartile on the OE test(s), in many cases in the bottom 10% of ranked performance, confirmed *twice*. Some differences are marked – the Asian participant flagged with an asterisk in Figure 3 (b9aX196i) performs in the 93rd percentile on the Asia CFMT, but only on the 6th for the Boston and 5th for the Australia CFMTs. The Caucasian sample is also shown in Figure 3 (bottom half), with a horizontal line demarcating performance on the measure of other (Asian) ethnicity consistent with performance in the bottom 10% of the population. The Caucasian participant marked with an asterisk (Zh177V0f)) scores in the top quartile for both Caucasian CFMT tests - 94% (Boston) and 78% (Australian) - but in the *bottom 8%* on the Asia CFMT.

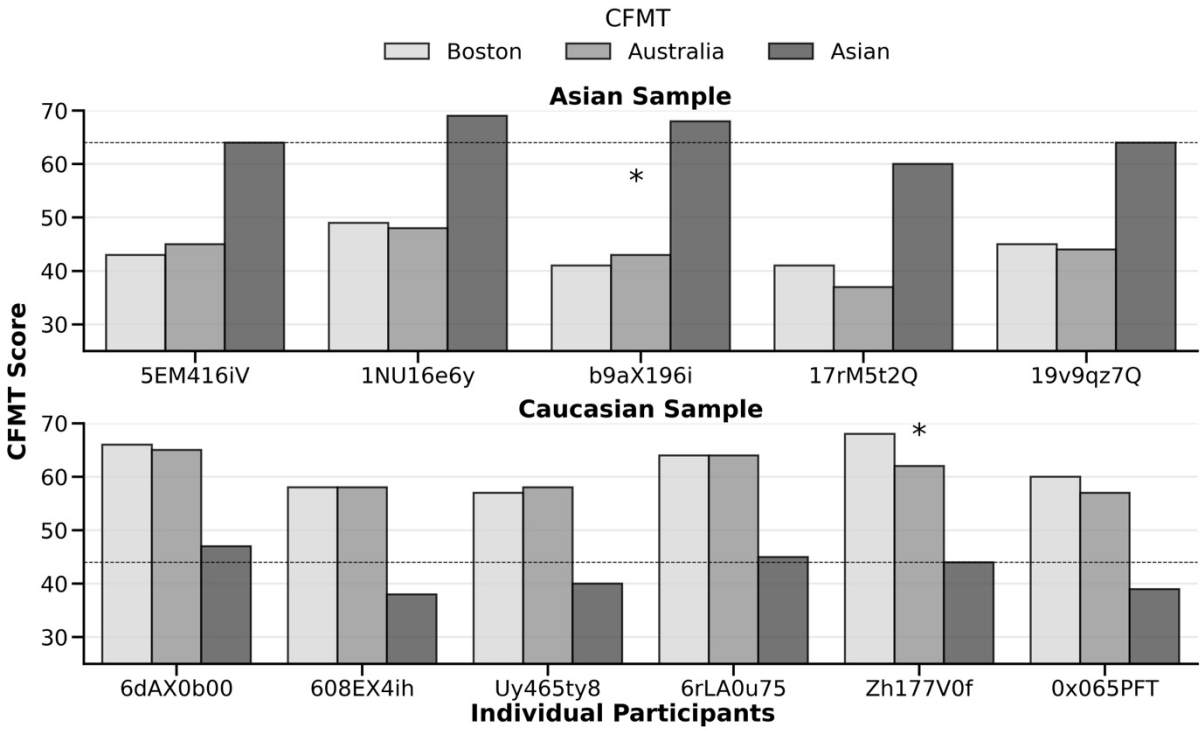

Figure 3 - Individual scores for candidate participant 'Hyper' OEE individuals – Note: top horizontal line marks performance that ranks in the top 15% of the general population for recognizing faces of the same ethnicity. Botton horizontal line marks performance in the bottom 10% of the population for recognising faces of the different ethnicity. Individual cases marked with an asterisk are discussed in the text.

In both cases, it is clear there are individuals who show very large and statistically significant disparities across own/other ethnicity testing that survives multiple comparisons. However, because their performance is not very low on the other ethnicity test, these individuals are missed by the 2% cut off criteria imposed by our initial approaches. In effect these individuals are instances of the opposite side of the coin to that in our discussion when using Wan and colleagues' percentile 2% cut-off method – since this typically identified generally low performers. In this case, because we have abandoned the focus on the very lowest other ethnicity face performance, we reveal that there are in fact individuals who appear to show striking performance differences across face types. Such a pattern has not been explored

before, since OEE studies describe differences at the average level. Our analyses suggest then that there are individuals who can show what is akin to a *hyper*-OEE, and we would suggest this likely warrants further investigation.

In summary, under this third approach, we sought out individuals with a performance difference across own/other ethnicity testing which is so sizeable that it is statistically significant. Using the Crawford RSDT approach, and considering *two* separate measurement comparisons, it appears that some individuals, though rare, can be found across both populations. In fact, in total we have identified 11/824 such candidates or 1.33% of the total sample.

Model-based inferences and posterior-predictive simulations

Thus far our analyses indicate - using multiple approaches and recognising regression toward the mean - that the rate of individuals with a notable performance difference between own and other ethnicity testing is small. However, we wish to extend these findings and overcome the three limitations described earlier - namely that uncertainty in the sample parameters impacts the classification of individuals, that there is a lack of a formal statistical model of how the sample data was generated, and finally, our inferences are tied to the sample data we collected. We describe here the results of a Bayesian posterior predictive model of the data.

Our approach proceeds in two stages. First, we use Bayesian inference to fit a multivariate-normal distribution model to the Caucasian and Asian datasets, separately. This distribution is ideal as a model, as it is parameterised by a vector of means and standard deviations (one for each of the three face recognition tests), and a correlation matrix, representing the association amongst measures - the precise estimates needed for the single-case tests. As this process is Bayesian, we obtain uncertainty in each of these estimates as probability distributions over each of them - for example, we obtain uncertainty in the correlation between the Boston and Australia CFMT's for the Asian sample, the mean of the Boston CFMT

for the Caucasians, and so on. We use relatively uninformative priors for these estimates - a normal distribution (with a mean of 50 and an SD of 10) for the means of each CFMT, a half-normal distribution with an SD of 10 for the standard deviations, and an LKJ prior with an eta of 1 for the correlations (which represents a prior belief any correlation between ±1 is possible). While the obtained *averages* of these posterior distributions of these estimates closely match those reported values in Table 2, for each of the CFMTs for each sample, the key additional information is the uncertainty in these estimates, which is central in the second stage of the analysis. We estimated these parameters using Markov-Chain Monte Carlo methods in the PyMC package of the Python programming language, which yielded 4,000 samples from the posterior distribution.

Using the posterior distribution of the estimates, we can obtain a piece of key inferential information - the *posterior predictive distribution.* For every sample obtained from the posterior (i.e., an estimate of the means, standard deviations, and correlations amongst measures), a multivariate normal distribution is instantiated, and 10,000 observations were drawn from it. This effectively creates new, simulated datasets of a very large sample size from the model, whilst propagating the uncertainty in the parameter estimates into the new data (because the procedure is repeated for each of the 4,000 posterior samples). This was done for both the Caucasian and Asian samples, resulting in 4,000 new datasets per ethnicity, each with 10,000 observations, and with means, standard deviations, and correlations similar to the initially collected data. This posterior predictive distribution allows us to address the three main limitations of our initial analysis - we propagate uncertainty in the parameter estimates used for the single case tests, generating plausible datasets for every plausible mean, standard deviation, and correlation matrix *given data we collected*. We do this by specifying a formal statistical model that generates new data that mimics the original data, and we can increase the sample size of this new data to a large value.

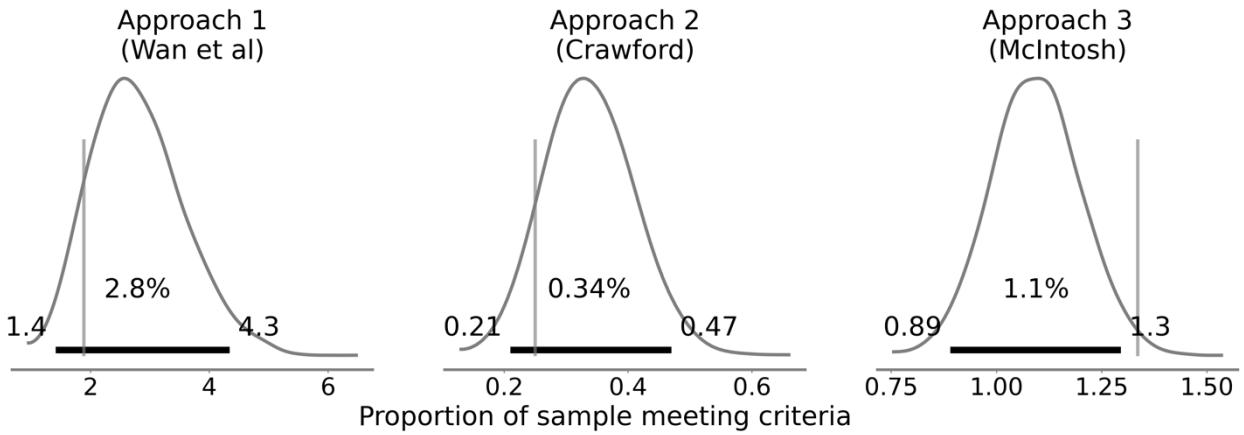

Figure 4 - Distributions of rates of classified individuals across the three approaches, obtained from a posterior predictive simulation. For the Wan et al and Crawford approaches, rates represent observed frequency out of 10,000 individuals (Asian and Caucasian, respectively), while for the McIntosh approach, the rate is frequency out of 20,000 (half of each ethnicity). Vertical lines represent the actual rates observed in the sample data.

Using this large collection of simulated datasets, we repeated our single-case tests on each, using the same three approaches as described earlier, extracting from each pair of Caucasian and Asian datasets the proportion of participants fitting the criteria. This yielded a probability distribution of rates for each approach, allowing us to describe the likely proportion of individuals who may be OEE-blind. For the approach used by Wan et al, we observed on average, 2.8% of Asian participants meeting the criteria, with a 95% credible interval between 1.3% and 4.2%. That is, out of a sample of 10,000 Asian individuals, somewhere between 130 to 430 individuals are a likely number of cases given that approach. The Crawford approach yielded a rate of just 0.34% [0.19, 0.46], less than one percent, suggesting that out of 10,000 Caucasians, between 19 and 46 individuals meet these criteria. Finally, for the McIntosh approach, a likely rate of 1.1% [0.89, 1.3] was observed across the entire samples (e.g., out of 20,000, obtained by combining each Caucasian and Asian sample together on each iteration), suggesting between 178 and 260 individuals would show an extreme pattern of dissociation.

These distributions are illustrated in Figure 4 above, with reference lines (vertically presented) illustrating the observed rate calculated from the original sample data. What is clear then is that our earlier estimates are indeed plausible for what we may observe in future data collection, and importantly demonstrates that in all cases the likely count is not *zero*.

Measurement error, regression to mean and the dangers of 'spurious' dissociations

Finally, we briefly revisit the concept of measurement error and regression to the mean here. Namely that a single score on two tests to be compared may reflect unusually high/low performance in the range for those scores a given participant might generate on these tests - which makes interpreting just *one* pairwise observation (e.g., Boston/Asia) potentially problematic. As was mentioned earlier, in our Wan and colleagues analyses we did identify some potential candidate Asian individuals who met criteria for a OEB 'dissociation' pattern, on one observed pairwise comparison (i.e., Boston/Asia or Australian/Asia – see Appendix 2), but this pattern didn't remain *twice* – in fact performance for many such individuals at second observation was often entirely normal. Thus, if these singular pairwise observations reflect 'error' in our measurements, we might expect to see instances in our sample of *spurious* 'dissociative' patterns, such as participants who score below a cut off *on their own ethnicity test* and yet appear to do much better on a comparative *other ethnicity* test. Upon investigation, across the entire sample we did find 6/824 such individuals (see Table 5 below) who appeared to show such a 'spurious' dissociation - that is they appeared to have a category specific *own ethnicity* deficit. Put simply, in a specific own/other ethnicity pairwise comparison we appear to have two Caucasians and four Asians who are performing at a level of being classified *developmental prosopagnosic* on their own ethnicity measure (scores marked in bold) and significantly *better* on a measure of other ethnicity.

Importantly however, this dissociative pattern *did not* persist across multiple comparisons. If we consider the example of the two Caucasian individuals - their apparently

poor own ethnicity performance on the CFMT Boston did not replicate on our second measure of own ethnicity performance (Australian, see Table 5). In the case of the three Asian individuals, two showed a very large inverted OEE for one comparison but not the other. These investigations clearly highlight the dangers of interpretation across just *one* pair of measures in the identification of OEB candidates (as was the case in the work of Wan and colleagues) - since in the examples mentioned you have individuals who appear to show a nonsensical dissociation based only on a comparison of two values, where one score likely doesn't reflect the participant's more general performance ('true' score) - that is, there are some singular examples which are likely simply 'error' in our measurements and thus do not survive repeated observations due to measurement error and RTM. In sum, the message is simple: we *must* use multiple pair-wise comparisons to have confidence in any 'abnormal' patterns we might be observing (see Murray & Bate, 2020 and Degutis et al., 2022, for similar commentaries regarding the classification of developmental prosopagnosia) - a lesson all too familiar to those working in the tradition of classical neuropsychology.

Participant ID	*Participant Country*	*Age*	*Gender*	*Bos Total*	*Aus Total*	*Asia Total*
A1dV301X	UK	19	Woman	**37**	49	54
x9749AgO	Serbia	18	Man	**40**	58	55
G82alo41	Singapore	19	Woman	52	46	37
B340xG4k	SKorea	21	Man	43	57	**37**
43H4Kb5v	Singapore	19	Woman	41	50	**38**
0sLA1w43	Japan	20	Man	51	42	**39**

Table 5 - Candidates for 'spurious' or counter-intuitive inverted OEE dissociations.

General Discussion:

The current work sought to investigate the prevalence rates of a presentation dubbed 'other ethnicity blindness' (OEB) in some large samples of Asian and Caucasian populations. In general, the research that has considered the other-ethnicity effect (OEE) has largely focused on the population level, but it is apparent that *within* a population individuals vary substantially in their face recognition abilities across both face category types - and the current work speaks to this individual differences issue. In addition, when considering the prospect of identifying individuals at 'extremes' of a distribution of performance, our work served to stress the relevance of measurement error and the related issue of regression to the mean (RTM) by using multiple observations. Since we have illustrated that, without taking such precautions, there is a danger that the identified prevalence rate may be much higher than is realistically the case.

Overall, adopting this multiple comparison approach we can provide prevalence estimates based on three different analyses:

(a) Using a percentile rank 'cut-offs' approach (suggested by Wan and colleagues), confirmed across *two* pairwise comparisons (i.e., Boston/Asia and Australian/Asia), we found that 1.9% (N=8/424) of the Asian population met criteria, much lower than the 8.1% prevalence described by Wan et al (2017). We further explored the plausibility of our earlier estimate, by using a Bayes analysis approach, under which we observed on average, 2.8% of Asian participants meeting the criteria, with a 95% credible interval between 1.3% and 4.2%. Importantly, closer inspection of the 8 individuals so classified from our sample - determined that nearly all performed poorly with *own-ethnicity* faces, with only a *single* individual scoring at or above average with these items (5iv74HR7). Interestingly, Wan and colleagues found a very similar pattern with few (near zero) OEB candidates performing at/above average with own-ethnicity

faces (see Figure 1). As a consequence, we would argue that OEB so classified is likely largely characterised by quite poor *general* face ability.

(b) Having established under this earlier approach that both our work and that of Wan and colleagues observed a very small number of individuals who have a pattern of OEB which cannot be dismissed as generally poor performance, we adopted a second *single case* 'dissociation' approach, drawing on Crawford's statistical tests. We again found only a *single* individual (1/400 - 0.25%) in the Caucasian sample, who was impaired with other ethnicity faces and *significantly better* with own-ethnicity faces, confirmed across *two* pairwise comparisons - and thus unlikely to be explained by generally poor face performance. On balance, if we define OEB as a form of genuine "blindness" for other-ethnicity faces in a manner akin to a *category specific* 'dissociation', then few individuals meet such criteria. But this observed number is *not zero* (likely 0.34% [0.19, 0.46] based on our Bayesian analysis). This may well warrant further investigation, since little is known of within category 'dissociative' patterns in the general population.

(c) Our third approach sought to identify individuals in the sample presenting with a *'hyper' OEE*, defined as a within-individual statistically significant difference across face categories (using RSDT) - which reflects an OEE x3-x4 times greater than the general average, again confirmed *twice*. This approach was inspired by the work of McIntosh (2018) who argued key dissociations could more simply be classified by such within individual disproportionately different performance. Overall, we find that 1.33% (11/824) of participants across our entire samples combined appeared to show this 'hyper' OE effect, that is a *disproportionately large difference across own and other ethnicity faces* (regardless of *absolute performance* on one or other). We confirmed the plausibility of this prevalence estimate with our subsequent Bayes analysis which suggested a likely rate of 1.1% [0.89, 1.3]. This pattern has not been explored before, since

OEE studies tend to describe the pattern at a macro level (i.e., population averages). We would argue that this approach is likely to be the most fruitful in terms of identifying OEB candidates - in that, we would re-define "OEB" as *individuals manifesting an abnormally large OEE.* With OEE being the relative difference between own/other ethnicity face recognition memory measured *twice* (to reflect the consequences of measurement error and RTM). On balance, we would argue that individuals manifesting such a pattern (if confirmed) would be intriguing since such 'hyper-OEE' performance would have both potential real-world consequences (e.g., forensic situations) and implications for cognitive models of face recognition memory which we touch upon in the earlier discussion and will return to in our implications section below.

Although we have indicated that the seminal work of Wan and colleagues could not adopt our approach, as they did not have a further measure to use, Wan and colleagues did attempt to address the issue of RTM by accounting for the correlation between measures. Using this approach, they reported 11/444 (2.5%) of their total remained consistent with an OE blindness performance pattern (8 Caucasian and 3 Asian participants, see the dashed circles in Figure 1 presented in the Introduction). We would thus argue this analysis aligns with our own findings (our Bayesian analysis observed a credible interval of 1.3-4.3%), though we would still urge the use of multiple measures rather than statistical 'correction'. Thus, on balance, although our findings appear not to agree with the 'headline' rate of OEB reported by Wan and colleagues (i.e., 8.1%), we do align with their estimates that attempted to account for RTM and feel this is a much more appropriate 'headline' to report.

Beyond addressing the above issues, through applying Bayesian inference, we addressed further issues with the single-case approaches. We specified a statistical model of the data, which was that the observed scores across the CFMT's of any ethnicity, in either sample, were generated by a multivariate normal distribution. Starting on this assumption, we used Bayesian inference to estimate the uncertainty in the means, standard deviations, and correlations of this distribution, and generated plausible datasets across these uncertainties.

This analysis confirmed several things - first, that this model of the data is a plausible, as it generated datasets in line with what we observed - all of the OEB rates in our actual sample fell within the distributions predicted by the model. Second, this approach suggests that the rate of OEB under any of the suggested 'classifications' is certainly *nonzero* (that is, some individuals in a large enough population will be observed), but it is also likely very small. Usefully, by simulating large populations of individuals (i.e., samples of 10,000 observations), we found that the highest plausible value was around 450 individuals (using the Wan et al approach), with most estimates being far smaller. As such, we show clear evidence OEB - rigorously classified - is likely very rare but *does plausibly exist*.

The current work provides a test-case of adopting a *different approach* to identifying individuals at 'extremes' of performance in this context, inspired by the work of McIntosh (2018). We sought to determine the prevalence of individuals that demonstrate an 'extreme' performance pattern that relates solely to a *relative difference* across categories. Considering OEB from this perspective raises important questions. For example, why might someone who can perform in the *93rd percentile* with own ethnicity faces appear to be so *consistently* poor (6th and 5th percentile) with other ethnicity faces? It is this *disproportionality* that raises such intriguing questions, rather than observing someone who appears singularly poor with a single category (which is a disadvantage of the Wan approach discussed earlier). This approach also side-steps the issues of computing and using 'cut-offs' as inclusion criteria to quantify impaired ability, with the potential problems that such hard boundaries generate when interpreting individual category performance - notwithstanding the related issue of proving a 'null' hypothesis for the 'preserved' task performance, and the fact that each category cut-off comparison statistically treats the observations as *independent* (even though they come from the same individual) - in addition this approach draws on reports of observed *correlations* between tasks to further highlight the extreme nature of performance disparity. The key examples observed

here would have been *entirely ignored* if the 2% cut off criteria to classify impaired other ethnicity performance was employed.

A reviewer helpfully pointed out that part of the reason why RTM may be so important in this case is that CFMT test-retest reliability may be subpar, that is it reflects issues of measurement error. As was mentioned both in the earlier work of Wan and colleagues and our own, the internal reliability correlations for our CFMTs were all consistently high (.85+), but it has been previously reported that test-retest reliability for at least one of the CFMT used (CFMT-Boston) was as low as .68 (Murray & Bate, 2020), although the 'upper bound' would be much higher. We entirely agree with this point, and in fact would suggest that it also demonstrates why our suggestion for using the McIntosh approach of focusing on extreme *differences* across tests has such utility. Clearly, measurement error has important implications both with respect to the observation of a *single score* to interpret 'true' within subject performance (as we have argued), but in the case of two observed scores appearing extremely different the strong likelihood is that because of RTM repeated observation of the same profile is *even more unlikely* – X *and* Y should converge on second observation if measurement error is driving observations. This is precisely why our evidence showed 'spurious' observations uniformly vanished in the same scenario. It is certainly possible that some OEB cases may have slipped from view as the reviewer suggests but reiterate the issue they have raised actually also implies that the observation of 'survivor' candidates is all the more striking. In any case, we would agree that measurement error *may* mean our observations are conservative, but our Bayesian analysis speak to this issue and we suspect the ranges we suggest remain highly plausible.

In fact, following this suggestion, it is perhaps important to stress that the initial impetus for using Z-scores (or percentile ranks) as 'cut-offs' for classifying abnormal individual performance, draws from the field of classical cognitive neuropsychology - where patient performance is typically clearly impaired, as individuals have experienced some form of brain

injury. As such, the observed patterns of performance are expected to be *qualitatively* different from that of the general population (i.e., very large z-scores), in that a patient was otherwise normal before their brain injury, and the event has 'subtracted' from their functional system (Caramazza, 1984 - the *subtractivity assumption* see Saffran, 1982), with the empirical objective being to understand what this implies to our cognitive models of this process and avenues for rehabilitation (e.g., Ball et al., 2004, Code et al. 2006, Code, Tree & Dawe, 2009, Tree et al. 2001). However, in this OEB case the same is not true - there is no expectation that individuals who may be OEB to be like brain injury populations, and thus one *would* expect any individual pattern of performance to *be in the observed distribution* of the general population (see DeGutis et al, 2022, who make a similar observation about developmental prosopagnosia, arguing they are also not qualitatively different from the general population). Our results here suggest it may be more fruitful to consider extreme performance disparities within individuals *across* these distributions, as opposed to focusing on singular impairment 'cut-offs', since the aims of face processing researchers are not the same as that of classical cognitive neuropsychology.

What are the implications for the presence of 'OEE' blind participants if they exist?

The current work sought to challenge the initial reports of OEE 'blindness' first described in the seminal work of Wan and colleagues, by determining the degree to which such cases are observed across a variety of classification approaches, which all attempt to grapple with issues around measurement error and RTM. We conclude that prevalence rates are likely much lower, but that evidence remains that *some* individuals can indeed present with a striking dissociation across own/other ethnicity unfamiliar face recognition. Moreover, we provide the suggestion that OEB cases should be best identified solely on the basis of extreme dissociation or 'hyper' OEE and it is worth reflecting on the implications of OEB on our understanding of variability in human

face recognition. As we discussed in the earlier introduction, we would suggest there are two key themes, one practical and one experimental/theoretical.

In the first case, the presence of individuals who can perform as poorly with other ethnicity faces as OEB cases clearly do, despite appearing 'normal' with *own* ethnicity faces has obvious implications for their practical credibility as eyewitnesses. Thus, our work confirms that in legal contexts at least, assessing an eyewitness's likelihood of making an other-race misidentification necessitates understanding the witness's fundamental face recognition capabilities. Moreover, drawing a parallel with prosopagnosia (the general inability to recognize faces) would suggest that being OEB could also significantly affect every day social interactions involving other-ethnicity individuals, such as those among colleagues in professional environments – resulting in both misunderstood distress and anxiety for those in question. We hope that the current work can act as a spur to research that can highlight these 'real world' impacts in greater detail.

In a similar vein, a reviewer pointed out that a key related issue is the degree to which OEB individuals have *insight* into their within category performance 'dissociation'. In the field of 'extreme' individual differences and face recognition ability there is a lively debate around this issue, in that many individuals who 'self-identify' as extremely poor at everyday face recognition (DP) are often able to perform at levels above CFMT impairment 'cut-offs' (see Burns et al., 2023 for a compelling discussion). At the same time a great deal of research on subjective awareness of one's own cognitive performance (face processing or otherwise) suggests that observed population correlations between objective/subjective measures are often quite low (Bowles et al., 2009. Palermo et al., 2016) – which leaves the challenge of how best to interpret this pattern (see Kramer & Tree, 2024 for a discussion). We would be cautious to enter into this debate in this context but would agree with the reviewer that an obvious and important next step to understanding OEB is also to explore the *insight* question – more than two decades ago, when researchers first started reporting their observations of DP it is fair to say the field was

initially very sceptical. Fast forward to the present day, and DP awareness both in the research field and the general public is much greater (e.g., the NHS recognises the condition) – this awareness likely also acts as an important spur to the recruitment of candidates who 'self-report' as having every-day problems and as we have mentioned the recognition of various 'real world' social/legal consequences – interestingly in the case of DP there have been instances were individuals identified via large scale 'screening', reported they had no prior awareness that it was a face recognition problem that was the root cause of their social interaction challenges (Susilo et al., 2010), so one can imagine the same is true for OEB. It would therefore be interesting to see if our efforts into the nascent observations of OEB can act as a similar spur, and hope the reviewer feels similarly inspired to explore this further.

In the second case, the observation of individuals who can manifest a *within* visual category 'extreme dissociation' is (as far as we are aware) beyond the scope of *any* theoretical models of face processing (or in fact visual learning more generally). How best to explain the observation of such extreme dissociations is a future challenge for *all face processing models* assuming that they aim to capture the diversity of human experience in recognising faces. We believe that it is in fact this remarkable range in observed human performance that makes psychology so fascinating a research area; mere focus on the 'average' misses these issues - the fact that such a considerable diversity of performance occurs (namely that there are individuals who show an apparent within category dissociation) on a task as fundamental to the human experience as face processing is we believe worthy of further investigation. The diversity on other processes of human cognition (e.g., short-term memory, 'inhibitory' processing, aspects of visual attention) are thus very likely to demonstrate similar striking dissociations, and yet this is largely ignored by the field making any such work simply incomplete (although see examples of lively discussion with respect to individual differences in 'inhibitory' processing – e.g., Hedge, Powell & Sumner, 2018, Rouder, Kumar, Haaf, 2023).

Moreover, theories of face processing that have provided an account for the generally impaired pattern of performance seen in DP, have suggested that this reflects disruption to particular perceptual mechanisms argued to be heavily utilised in this case. For instance, individuals with prosopagnosia may show deficits in 'holistic' processing, such as inversion effects (Farah, Wilson, Drain, & Tanaka, 1995), sensitivity to the spacing between facial features (Yovel & Duchaine, 2006) and impairments of face-space coding (Palermo et al., 2011). In a similar vein, other theoretical accounts have argued that the OEE arises because of differential deployment of 'configural' (own) versus 'featural' (other) ethnicity face performance (e.g., Esins et al., 2014; Michel et al., 2006; Mondloch et al., 2010). However, as Wan and colleagues pointed out, all these studies focus on the *average effects* observed, rather than focusing on 'extreme' (OEB) cases. As a consequence, such accounts have not specified the extent to which this difference may extend - the assumption is largely that there may be some decrease in performance but not to the degree we observed in certain cases. In addition, a further question arises as to whether this within category 'dissociation' observed in OEB is *face specific* or whether it may be underpinned by some more general difficulty with making certain kinds of *within* class discriminations – which speaks to theories of 'face specificity' which remain an ongoing debate. Overall, we would say simply that if the objective of science is to account for the full range of observed data for any topic under scrutiny; ignoring the extent of distributions is simply not good science.

Related to the interpretation of the OEB presentation performance, a reviewer made the important suggestion that it is also possible that extreme performance examples we have observed may manifest for reasons outside face processing per se, occurring because of individual differences linked to social or motivational factors. For example, social-motivation theories of the OEE might suggest that poor recognition of other-race faces is due to lack of effort applied to individuating other-race people (Hugenberg et al., 2007; MacLin & Malpass), and is in fact an issue that was considered in the original work of Wan and colleagues. In their

case they found no evidence for such effects, in fact the general pattern was for the *opposite* in that OEB candidates reported putting *more* and not less effort into the tasks. In earlier work, Wan et al. (2015) argued that the likely drivers for previously observed social-motivation contributions to the OEE are likely to be linked to situations across high and low socioeconomic status (typically US Whites vs US Blacks), which wouldn't likely apply to the samples we have observed here given the different cultures represented (similarly for the earlier Wan et al., 2017 study). In any case, we agree with the reviewer in that this remains an interesting point and warrants exploration in future with individuals identified using the methods suggested here. Moreover, if it were indeed the case that such factors could manifest such extreme performance, the implications are also important – since just as it is true that no cognitive models of face processing predicts (or even recognises) that such extreme *within* category performance may manifest, we are not aware that experimentalists might recognise that motivation could have such an extreme consequence. In either case, empirical questions are left begging, and the impact of ignoring such individual variance has practical consequences on data collection and analysis.

In conclusion, the current work has provided a clear roadmap as to how one may appropriately identify individuals who may be candidates for OEE 'blindness', that is individuals who manifest extremely poor face recognition ability with faces other than their own ethnicity. We hope that the various potential avenues for future research both practical and theoretical may be considered using this roadmap we have provided and that more consideration of individual difference 'dissociations' in cognition will follow.

Constraints on Generality:

The findings of this study are drawn from a comparatively large opportunity sample of young adult participants from Asian (Japan, Korea, China, and Singapore) and Caucasian (Australia, the UK, and Serbia) backgrounds, recruited through an online platform. While the sample commendably includes diverse cultural backgrounds, results may not generalize to older adults, children, or individuals from other ethnic groups. Additionally, all face recognition assessments were conducted using the Cambridge Face Memory Test (CFMT) in its Boston, Australian, and Asian versions, which are validated for measuring face recognition ability but may not fully capture the complexities of real-world face recognition. As a consequence, some of the prevalence figures we present may be impacted by natural measurement error, although our simulation analysis approach attempts to address this issue. The online nature of the study could introduce variability in test conditions (e.g., device type or screen resolution), though our software's design aimed to standardize participant experience.

Public Significance Statement:

Our research highlights a group of individuals who experience a severe impairment in recognizing faces from other ethnicity groups, a condition akin to "face-blindness" (prosopagnosia) *specifically* for other-race faces. These findings provide new insight into the "other-race ethnicity effect" (OEE), which describes how people often find it harder to recognize faces from racial groups different from their own. While the average effect of OEE may seem modest, this research shows that some individuals experience significant challenges, which can have serious real-world consequences. In legal contexts, for instance, these difficulties in recognizing other-ethnicity faces may lead to mistaken eyewitness identifications, potentially resulting in wrongful convictions. Moreover, these findings suggest that in everyday social and workplace interactions, some individuals may struggle significantly with recognizing colleagues from different racial backgrounds, similar to how people with prosopagnosia struggle with all

faces. Understanding these specific challenges can inform policies in both legal and social settings to reduce bias and improve cross-racial interactions. Furthermore, this work sheds light on the broader diversity of individuals in the levels of their face recognition abilities, emphasizing the importance of considering individual differences to deepen our understanding of human cognition.

Open Science Statement

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. Additionally, the datasets supporting the conclusions of this article are available in the Open Science Framework repository, accessible at https://osf.io/yfgvb/

Materials and Methods Transparency

The materials and methods used in this research are comprehensively described in the Methods section of the article.

Code Availability

The custom scripts and algorithms used for data analysis are openly available and can be accessed via the Open Science Framework at https://osf.io/yfgvb/. The code is released under the CC-BY license permitting free use, modification, and distribution.

Ethical Compliance

All experiments and procedures were conducted in compliance with ethical guidelines and approved by the Swansea University Ethics board. Informed consent was obtained from all participants, and their privacy rights are respected.

By committing to these open science practices, we aim to enhance the transparency, reproducibility, and collaborative potential of our research. We welcome feedback and collaboration from the scientific community to further validate and extend our findings.

Declaration Statement

Conflicts of interest/ Competing interests – there are no conflict of interest/competing interests relating to this manuscript

Consent for publication – the authors consent for the publication of this manuscript if accepted.

References:

Ball, M. J., Code, C., Tree, J., Dawe, K., & Kay, J. (2004). Phonetic and phonological analysis of progressive speech degeneration: A case study. *Clinical linguistics & phonetics*, *18*(6-8), 447-462.

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.

Bate, S., Bennetts, R.J., Gregory, N., Tree, J.J., Murray, E., Adams, A., Bobak, A.K., Penton, T., Yang, T., Banissy, M.J., (2019) Objective patterns of face recognition deficits in 165 adults with self-reported developmental prosopagnosia. *Brain Sciences*, 9, 133.

Bennetts, R.J., Gregory, N., Tree, J.J., Banissy, M., Murray, E., Adams, A., Penton, T. & Bates. S. (2022). Featural and holistic processing can be separably impaired in disorders of face recognition Evidence for two subtypes of developmental prosopagnosia. *Neuropsychologia*, 174, 108332.

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, 30(1), 81-91.

Burns, E. J., Gaunt, E., Kidane, B., Hunter, L., & Pulford, J. (2023). A new approach to diagnosing and researching developmental prosopagnosia: Excluded cases are impaired too. Behavior research methods, 55(8), 4291-4314.

Campbell, D. T., & Kenny, D. A. (1999). A primer on regression artifacts. *Guilford Press.*

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and language*, *21*(1), 9-20.

Code, C., Muller, N., Tree, J., & Ball, M. (2006). Syntactic impairments can emerge later: Progressive agrammatic agraphia and syntactic comprehension impairment. *Aphasiology*, *20*(9), 1035-1058.

Code, C., Tree, J. J., & Dawe, K. (2009). Opportunities to say 'yes': rare speech automatisms in a case of progressive nonfluent aphasia and apraxia. *Neurocase*, *15*(6), 445-458.

Chen, W., Kassa, M.T., & Cheung, O.S. (2023). The role of implicit social bias on holistic processing of out-group faces. *Cognitive Research: Principles and Implications*, 8, 7.

Cheung, O.S., Quimpo, N.J., & Smoley, J. (2024). Implicit bias and experience influence overall but not relative trustworthiness judgment of other-race faces. *Scientific Reports*, 14, 16068.

Childs, M. J., Jones, A., Thwaites, P., Zdravković, S., Thorley, C., Suzuki, A., ... & Tree, J. J. (2021). Do individual differences in face recognition ability moderate the other ethnicity effect?. *Journal of Experimental Psychology: human perception and performance*, 47(7), 893.

Crawford, J. R., Garthwaite, P. H., & Gray, C. D. (2003). Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, 39(2), 357-370.

Crawford, J. R., Howell, D. C., & Garthwaite, P. H. (1998). Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples t test. *Journal of Clinical and Experimental Neuropsychology*, 20, 898-905.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist*, 23, 1173-1195.

Crawford, J. R., Garthwaite, P. H., and Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. Cognitive Neuropsychology, 27, 245-260.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology*, 30(6), 827-840

DeGutis, J., Bahierathan, K., Barahona, K., Lee, E., Evans, T. C., Shin, H. M., ... & Wilmer, J. B. (2022). What is the prevalence of prosopagnosia? An empirical assessment of different diagnostic cutoffs.

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.

Esins, J., Schultz, J., Wallraven, C., & Bülthoff, I. (2014). Do congenital prosopagnosia and the other-race effect affect the same face recognition mechanisms? *Frontiers in Human Neuroscience, 8,* 759.

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),

Fry, R., Wilmer, J., Xie, I., Verfaellie, M., & DeGutis, J. (2020). Evidence for normal novel object recognition abilities in developmental prosopagnosia. *Royal Society open science*, 7(9), 200988.

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

Gelman, A. Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733-807.

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.

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.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50, 1166–1186.

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,* 334 –340.

Jansari, A., Miller, S., Pearce, L., Cobb, S., Sagiv, N., Williams, A. L., ... & Hanley, J. R. (2015). The man who mistook his neuropsychologist for a popstar: When configural processing fails in acquired prosopagnosia. *Frontiers in Human Neuroscience*, *9*, 390.

Kennerknecht, I., Grueter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., & Grueter, M. (2006). First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *American Journal of Medical Genetics Part A*, *140*(15), 1617-1622.

Kramer, R. S., & Tree, J. J. (2024). Investigating people's metacognitive insight into their own face abilities. *Quarterly Journal of Experimental Psychology*, *77*(10), 1949-1956.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206.

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law, 7,* 98 –118.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of personality and social psychology*, 13(4), 330-334.

McIntosh, R. D. (2018). Simple dissociations for a higher-powered neuropsychology. *Cortex*, 103, 256-265.

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.

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.

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.

Michel, C., Caldara, R., & Rossion, B. (2006). Same-race faces are perceived more holistically than other-race faces. *Visual Cognition, 14,* 55–73.

Mondloch, C. J., Elms, N., Maurer, D., Rhodes, G., Hayward, W. G., Tanaka, J. W., & Zhou, G. (2010). Processes underlying the cross-race effect: An investigation of holistic, featural, and relational processing of own-race versus other-race faces. *Perception, 39,* 1065–1085.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science*, *7*(9), 200884.

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.

Palermo, R., Rivolta, D., Wilson, C. E., & Jeffery, L. (2011). Adaptive face space coding in congenital prosopagnosia: Typical figural aftereffects but abnormal identity aftereffects. *Neuropsychologia, 49,* 3801–3812

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin and Review*, 30(6), 2049–2066.

Saffran, E. (1982). Neuropsychological approaches to the study of language. *British Journal of Psychology*, 73, 317–337.

Trawiński, T., Aslanian, A., & Cheung, O.S. (2021). The effect of implicit racial bias on recognition of other-race faces. *Cognitive Research: Principles and Implications,* 6, 67

Tree, J. J., Perfect, T. J., Hirsh, K. W., & Copstick, S. (2001). Deep dysphasic performance in non-fluent progressive aphasia: A case study. *Neurocase*, *7*(6), 473-488.

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.

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.

Yovel, G., & Duchaine, B. (2006). Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia. *Journal of Cognitive Neuroscience, 18,* 580–593.

Zhao, M., Hayward, W. G., & Bülthoff, I. (2014). Holistic processing, contact, and the other- race effect in face recognition. *Vision Research*, 105, 61-69.

Zhou, X., Elshiekh, A., & Moulson, M. C. (2019). Lifetime perceptual experience shapes face memory for own-and other-race faces. Visual Cognition, 1-14.

Data and code availability: https://osf.io/yfgvb/

Participant ID	*Country*	*Race*	*Age*	*Gender*	*Aus_Total*	*Asia_Total*
35wj7E6m	China	Asian	20	Male	31	39
zs7U366k	Singapore	Asian	22	Male	34	40
GX605o4a	China	Asian	19	Male	36	40
h42jFq83	China	Asian	19	Male	38	40
40E8r5HS	Skorea	Asian	20	Male	33	42
312lxn4e	Skorea	Asian	20	Female	34	42
69Hc31SP	Skorea	Asian	20	Female	37	44
3r75KI7Q	Japan	Asian	19	Female	38	45
g3203Wgh	Japan	Asian	18	Male	39	45
9MZd712y	China	Asian	18	Male	34	46
1yEn750C	Japan	Asian	19	Male	36	46
2Q6dc34I	China	Asian	18	Male	35	47
C0fP34b6	Japan	Asian	21	Female	35	47
B112Kv3u	China	Asian	20	Male	36	47
wp9Y89L2	Singapore	Asian	23	Male	39	47
lvZ047n5	Japan	Asian	20	Female	37	48
52FAV6d9	Singapore	Asian	21	Male	33	49
7f19Uh5u	Japan	Asian	19	Female	38	49
hQ656b7k	Japan	Asian	20	Male	39	49
F902EGM6	Singapore	Asian	22	Female	37	50
QD4h66d6	Japan	Asian	25	Male	39	51
vE7910Kk	Japan	Asian	18	Female	39	51
0Gu5KV09	Singapore	Asian	21	Male	37	52
04JrNR60	Skorea	Asian	21	Male	35	53
Q25o7eX8	Japan	Asian	21	Female	36	53
q171Yvh4	China	Asian	18	Female	37	54
84PFPf42	Japan	Asian	20	Male	38	54
P1U05X2F	China	Asian	19	Female	39	55
I2Sv9d84	Singapore	Asian	19	Female	37	56
q684hvl8	Skorea	Asian	20	Female	39	56
5iv74HR7	Japan	Asian	20	Male	39	57
T312UW8Z	Japan	Asian	19	Male	39	57
48rsFQ07	China	Asian	19	Male	37	59
6Ie8GA49	Skorea	Asian	20	Female	37	59
17rM5t2Q	Japan	Asian	19	Male	37	60
3m11X2vo	China	Asian	19	Female	37	63

eZG1i910	UK	Caucasian	18	Female	40	34
QPb925x7	Australia	Caucasian	26	Female	43	34
b7O1Jm71	Australia	Caucasian	18	Female	44	38
j7P4ts31	Australia	Caucasian	19	Female	44	38
65Kwzu80	UK	Caucasian	19	Male	45	38
UW8OP707	UK	Caucasian	22	Male	45	38
48rc5cK0	UK	Caucasian	18	Male	46	36
n56D6AP9	UK	Caucasian	18	Female	46	38
vr468Yj8	Serbia	Caucasian	20	Male	47	37
s1894mKG	Australia	Caucasian	19	Female	47	38
12ogv5i3	Serbia	Caucasian	19	Male	49	34
PZ3E39f7	UK	Caucasian	18	Female	50	33
E28sJ58u	Australia	Caucasian	18	Male	50	38
r3961wxP	Australia	Caucasian	20	Male	50	38
K379zA4h	Serbia	Caucasian	23	Male	50	38
65TE3H5p	Serbia	Caucasian	19	Female	51	38
K479Nc4g	Serbia	Caucasian	21	Female	53	37
M5PY265s	Serbia	Caucasian	21	Male	56	36
608EX4ih	Serbia	Caucasian	19	Female	58	38

Appendix 1 – Individuals who meet ‘cut-off’ on Asia or Australian CFMTs

Participant ID	*Country*	*Age*	*Gender*	*Boston Total*	*Australia Total*	*Asia Total*	*Impairment*
35wj7E6m	China	20	Male	**33**	**31**	39	1
h42jFq83	China	19	Male	**36**	**38**	40	1
40E8r5HS	Skorea	20	Male	**33**	**33**	42	1
1yEn750C	Japan	19	Male	**32**	**36**	46	1
9MZd712y	China	18	Male	**37**	**34**	46	1
B112Kv3u	China	20	Male	**35**	**36**	47	1
52FAV6d9	Singapore	21	Male	**30**	**33**	49	1
5iv74HR7	Japan	20	Male	**37**	**39**	57	1
Pp39T5O9	Singapore	19	Female	**37**	44	39	2
TP73Xf52	China	20	Male	**33**	44	40	2
044Fyc0O	China	18	Male	**37**	48	40	2
M1dT0Z90	Skorea	20	Male	**37**	50	40	2
g1x7JG35	Japan	19	Male	**34**	46	42	2
T0Jt902P	China	19	Male	**33**	41	44	2
64Jq78wd	China	20	Male	**35**	44	44	2
tC20tj23	China	19	Male	**36**	42	44	2
21f9L0Xi	China	18	Male	**33**	45	45	2
2oOmU120	China	18	Female	**33**	50	48	2
3G8Kt8K1	China	19	Male	**35**	45	48	2
85U8P3Xy	Singapore	20	Female	**37**	42	48	2
O2zY9R08	Japan	19	Male	**37**	47	48	2
vGF42Y53	China	19	Male	**34**	54	51	2
3M58faf9	China	19	Male	**37**	56	51	2
Wu1m06W2	Singapore	20	Female	**33**	45	52	2
3xkM82W3	Skorea	19	Male	**37**	44	52	2
D930R5Zy	China	18	Female	**34**	41	54	2
YT05g69l	China	20	Male	**34**	53	54	2
VDl6M320	China	20	Female	**35**	49	67	2
GX605o4a	China	19	Male	40	**36**	40	3
zs7U366k	Singapore	22	Male	45	**34**	40	3
312lxn4e	Skorea	20	Female	42	**34**	42	3
69Hc31SP	Skorea	20	Female	44	**37**	44	3
g3203Wgh	Japan	18	Male	48	**39**	45	3
3r75KI7Q	Japan	19	Female	54	**38**	45	3
wp9Y89L2	Singapore	23	Male	38	**39**	47	3
2Q6dc34I	China	18	Male	46	**35**	47	3
C0fP34b6	Japan	21	Female	47	**35**	47	3
IvZ047n5	Japan	20	Female	52	**37**	48	3
hQ656b7k	Japan	20	Male	42	**39**	49	3
7f19Uh5u	Japan	19	Female	44	**38**	49	3
F902EGM6	Singapore	22	Female	50	**37**	50	3

vE7910Kk	Japan	18	Female	43	**39**	51	3
QD4h66d6	Japan	25	Male	48	**39**	51	3
0Gu5KV09	Singapore	21	Male	42	**37**	52	3
04JrNR60	Skorea	21	Male	49	**35**	53	3
Q25o7eX8	Japan	21	Female	55	**36**	53	3
q171Yvh4	China	18	Female	39	**37**	54	3
84PFPf42	Japan	20	Male	58	**38**	54	3
P1U05X2F	China	19	Female	44	**39**	55	3
I2Sv9d84	Singapore	19	Female	40	**37**	56	3
q684hvI8	Skorea	20	Female	54	**39**	56	3
T312UW8Z	Japan	19	Male	54	**39**	57	3
48rsFQ07	China	19	Male	49	**37**	59	3
6Ie8GA49	Skorea	20	Female	62	**37**	59	3
17rM5t2Q	Japan	19	Male	41	**37**	60	3
3m11X2vo	China	19	Female	47	**37**	63	3

Impairment Key – 1= Boston/Aus both impaired, 2= Boston only impaired, 3= Aus only impaired.

Appendix 2 – Asians who meet 'cut-off' on Boston only, Australian only and both CFMTs