Who Is That? Perceptual Expertise on Other-Race Face Comparisons, Disguised Face Comparisons, and Face Memory

Amy N. Yates[1], Jacqueline G. Cavazos[2], Géraldine Jeckeln[3], Ying Hu[3], Eilidh Noyes[4], Carina A. Hahn[1], Alice J. O'Toole[3], and P. Jonathon Phillips[1]

[1] *National Institute of Standards and Technology*
[2] *University of California, Irvine*
[3] *The University of Texas at Dallas*
[4] *University of Huddersfield*

Abstract

Forensic facial professionals identify faces more accurately than untrained participants on tests using high quality images of faces. Whether this superiority holds in more challenging conditions is not known. Here, we measured performance for forensic facial professional groups (facial examiners and facial reviewers) and a group of untrained control participants (undergraduates). We tested performance in three challenging tasks: other-race face identification, disguised face identification, and memory for faces. We note that the administration of the other-race and disguise tests here did not allow forensic professionals access to the time and tools they typically use in casework. On the other-race face identification task, both groups of forensic professionals' accuracies did not exceed the accuracy of the control participants. Examiners were more accurate than controls on impersonation disguise, but were not consistently more accurate than controls on evasion disguise. On the Cambridge Face Memory Test (CFMT+), examiners' performance was marginally better than controls; and reviewers and controls performed equally well. We conclude that examiners' face identification superiority does not generalize completely to identification of other-race and disguised faces.

1 Introduction

Face identification is an integral part of law enforcement. Face identification judgments made by forensic facial examiners can be presented as expert evidence in legal proceedings due to their skill and training. These experts perform detailed and careful comparisons of face images to determine whether the same

person (or identity) appears in two or more images (*e.g.*, crime scene image and a mugshot). Despite the consequential nature of this role, until recently, remarkably little was known about the accuracy of these forensic facial professionals relative to other untrained humans [1, 2, 3, 4]. A meta-analysis of recent studies [4] indicates that forensic facial examiners are indeed more accurate than novices at face identification, in the conditions that have been tested, *e.g.*, short exposure, varying image quality, inverted images, images with body and clothing information.

Tests of forensic professionals' face identification abilities, however, have focused on performance in relatively controlled conditions, such as frontal facing. In 2015, White *et al.* [1] performed a battery of perceptual identity-matching tests on forensic examiners. That is, they investigated how accurate examiners were at determining whether two images were of the same person or not with time limited to 30 seconds or lower; conditions varied for each test, such as upright or inverted images, images with only body information, etc. There were three groups of participants: forensic examiners ($N = 27$), a group of technical experts and biometric system administrators (labeled as *controls* in the paper, $N = 14$), and undergraduate students ($N = 32$). This tests were administered with short stimulus exposure times (at most 30 sec.). Forensic examiners were more accurate than untrained observers with 30 seconds of exposure time. With 2 seconds of exposure time, forensic examiners were still better than undergraduate students but were equal to the control group. This study was the first to demonstrate the superiority of forensic facial examiners over untrained subjects. Due to the short exposure times, and the fact that examiners performed the task without access to the tools normally available in forensic examination (*e.g.*, digital enhancement), the results can be considered a lower bound estimate of forensic facial examiner performance.

In a more recent study [3], examiners' performance was tested under circumstances more similar to those in which they work. Specifically, forensic professionals had access to their tools and procedures, and had ample time to apply these to their face identification decisions. This type of test has been referred to in the literature as a "black-box" or "closed-box" test, because researchers do not have knowledge about how identification decisions are made. Participants from five groups were tested: professional forensic facial examiners ($N = 57$), professional forensic facial reviewers ($N = 30$) (*cf.*, [3]), "super-recognizers" [5, 6] ($N = 13$), professional fingerprint examiners ($N = 53$), and university students ($N = 31$). The task was to match the identity of faces in high quality frontal images that were chosen to be highly challenging based on previous human and machine studies. Face examiners, reviewers, and super-recognizers performed more accurately than fingerprint examiners and students, and fingerprint examiners performed more accurately than students. Four face identification algorithms developed between 2015 and 2017 performed the same image comparisons. Machine performance over the two-year span of algorithm development tracked the "expertise" level of the human participants (2015 algorithm [7] $\approx$ students; 2016 algorithm [8] $\approx$ fingerprint examiners; earlier 2017 algorithm [9] $\approx$ professional face reviewers; later 2017 algorithm [10] $\approx$ face

examiners). The best performance was achieved by combining the judgments of individual professional face examiners with those of the best algorithm [10]. This study was the first to measure the face identification accuracy of professional forensic facial examiners using their normal work procedures, and the first study to directly compare recent face recognition algorithms based on deep convolutional neural networks (DCNNs)[11] to professional face examiners, face reviewers, and super-recognisers.

The studies carried out to date provide a benchmark accuracy for forensic face identification professionals in ideal conditions. Unfortunately, the conditions under which forensic facial professionals operate are often "demanding", if we consider the possibility that conditions that are challenging for untrained individuals might be problematic for professionals as well. For example, the psychology literature has established that face recognition is prone to error when untrained people are asked to recognize "other-race" faces (*e.g.*, [12]) and disguised faces (*e.g.*, [13]). Both cases are relevant in law enforcement face identification scenarios. Beginning with the former, it is well-known that face recognition accuracy is better for faces of one's own race than for faces of another race—a phenomenon known as the cross-race effect (CRE) (*e.g.*, [12]). The CRE has been replicated in dozens of studies with a variety of methodologies, including perceptual [14, 15], neural [16, 17, 18], developmental [19, 20], memory [21, 22], and eyewitness memory tasks (*e.g.*, [23]).

Similar to humans, the performance of early face recognition algorithms also differs for faces of different races [24, 15]. In one study, algorithms showed a CRE such that the geographic origin of algorithms (East Asian versus Western countries) interacted with the race of faces tested (East Asian versus Caucasian) [15]. The overall accuracy of computer-based face recognition has increased substantially over the last decade. The performance of these newer algorithms has resulted in their widespread use both in mundane (*e.g.*, social media) and consequential (*e.g.*, passport security, law enforcement) applications. However, race bias remains a problem for current face recognition algorithms [25, 26, 27, 28, 29]), which are based on DCNNs (*cf.*, [30, 31, 32]). Therefore, the use of these algorithms amplifies societal concerns about law enforcement applications.

Although race bias in computer-based face recognition has been studied intensively in recent years, it is not known whether forensic facial examiners perform comparably for faces of different races. Cross-race face identification effects have been reported for super-recognisers, a group shown to be equal in accuracy to forensic facial examiners and reviewers [3], but these findings are not convergent. One study reported a standard CRE [33] for super-recognisers, whereas a second found that super-recognisers perform *more* accurately on other-race faces [34]. It is currently unknown how examiners' and reviewers' accuracies are affected by faces of different races. The first goal of the present study is to determine whether forensic facial examiners and reviewers perform comparably for faces of their own race and for faces of a different race. To measure this, we administered a test with face image pairs from [15]. This dataset has an equal number of Caucasian and East Asian face image pairs.

A second challenging case relevant for forensic facial examiners and reviewers is the task of identifying disguised faces. For untrained participants, disguise can have dramatic detrimental effects on face identification performance [13, 35, 36]. In the most comprehensive study to date, Noyes and Jenkins created the *Façade* database of realistic facial disguises [37]. People in the Façade database altered their facial appearance in two ways: a.) to look "as different as possible from themselves" (evasion disguise) and b.) to appear "as similar as possible to another specific person with a similar appearance" (impersonation disguise). Using a face identity-matching paradigm, participants unfamiliar with the disguised individuals were substantially less accurate at matching identity in disguised versus unaltered faces. Evasion disguise proved especially difficult. Notably, even people personally familiar with the disguised individuals performed poorly in matching the identity of evasion-disguised faces.

Disguised faces are also problematic for DCNNs [38]. Network accuracy for identifying disguised and undisguised faces from the Façade database mirrors human accuracy [38] (*cf.*, [13]). Attempts to "familiarize" the network with the identities being tested improved performance. Averaging DCNN-generated face representations enhanced the network's ability to group diverse images of identities together, thereby improving performance on impersonation disguise. An identity contrast learning algorithm, enhanced the network's ability to separate DCNN representations of different identities, thereby improving performance on evasion disguise. These types of manipulations might be useful in security applications that deal with disguised faces.

The ability of forensic facial examiners and reviewers to "see through facial disguise" to identify faces has not been tested. It is important to understand the effects of disguise in security and law enforcement, knowing disguise is problematic for both untrained people and generically-trained DCNNs. Therefore, the second goal of the present study was to compare professionals with untrained students on the task of face identification under disguise. To investigate his effect, we use the Façade database images.

The third goal of this study was to compare face identification performance for professionals and control participants in a task that involves memory. Forensic facial professionals are trained to perform face comparisons, with all relevant images available. Human face processing skills, however, are tapped most commonly in daily life to compare the memory of a face with a perceptually present face. As noted, in ideal cases examiners have been shown to have superior face matching ability [1, 3, 2]. It is of theoretical interest to know also whether the superior perceptual identity-matching skills of examiners generalize to a face memory task more like the one people do most commonly. To address this question, we tested examiners with the long form of the Cambridge Face Memory Test (CFMT+) [5]—a test widely used to separate people with superior face memory to those with from typical face memory. In this test, the identity comparison must occur between a perceptually present face and the representation of that face (and others) in memory.

In the next sections, we present three experiments in which we tested professional forensic facial examiners, reviewers, and untrained (Caucasian and East

Asian) students. The other-race and disguised face identification tests were perceptual matching tests conducted under laboratory style conditions. This type of test offers a first comparative look at professionals and untrained control participants on these challenging tasks. In casework, however, examiners perform identifications under less constrained conditions (*e.g.*, with access to tools and procedures, and with ample time to examine images). Previous work [1, 3] has indicated this may be a lower bound estimate.

2 Experiments

Participants completed three tests to examine: 1.) own- and other-race face identification, 2.) disguised face identification, and 3.) memory for faces. To compare face professionals to the general population, we recruited participants from four groups: forensic facial examiners, forensic facial reviewers, Caucasian undergraduate students, and East Asian undergraduate students. All but one forensic professional reported at least some Caucasian ancestry, and no forensic professional reported any Asian ancestry. Therefore, we were unable to recruit professionals of specific races. We begin with an overview of the participant groups and test administration procedures. Then we proceed with a description of the three experiments.

2.1 Participants and Test Administration

2.1.1 Forensic Facial Professional Testing

A total of 35 forensic facial professionals participated in this study. Data collection took place between 2017 and 2019. Participants were not compensated. Participants were required to be at least 18 years of age and have completed training as an examiner or reviewer or be employed as an examiner or reviewer. All requirements were self-reported. One participant was removed from the study due to familiarity with the stimuli. The analysis included 16 examiners (7 female) and 18 reviewers (10 female). Age was categorized into decade-wide bins, detailed in Appendix C. Examiner bins ranged from 18–29 to 50–59: (mode age bin 30–39), and reviewer age bins ranged from 30–39 to 50–59 (mode age bin 40–49). The National Institute of Standards and Technology (NIST) Research Protections Office reviewed the protocol for this project and determined it met the criteria for "exempt human subjects research" as defined in 15 CFR 27, the Common Rule for the Protection of Human Subjects. For logistical reasons, test administration differed for participants within the forensic professional group. Some professionals were tested remotely and some were tested in-person at the Face Identification Special Working Group (FISWG) meeting in October 2019. Except for three examiners, all participants completed all three tests. The three examiners who did not complete all three tests ran out of time and are included in the analysis for tests they completed, but not in the tests they did not.

Professional participants tested prior to May 2018 completed the task remotely. Researchers at NIST emailed task links to participants via Survey

Gizmo[1]. Participants were allowed to take the tests in any order but were asked to complete each test in a single session. Although remote participants had four weeks to complete the tests, timing constraints within each experiment were identical for all groups of participants (remote and in-person). Specifically, for the other-race and disguised face tests, each face pair was presented for 30 seconds. Response time was not limited, and no feedback was provided. Within each test, trial order and image position were fixed across participants. The standard procedures outlined in [5] were followed for the CFMT+. Additional details are provided in the method section of each experiment.

For facial professionals tested in person, NIST administered the three tests in a single, in-person session on a NIST laptop. The face tasks were followed by a demographic survey (via Shiny v1.3.2 [39]). The other-race test, disguise test, and CFMT+ were administered with PyschoPy v3.1.5 [40].

At the outset, we note that all but one of the professional participants reported at least some Caucasian ancestry (none reported East Asian ancestry). Therefore, a full-crossover design was not possible for the professional group. However, students of both Caucasian and East Asian ancestry participated and so provide a control on stimulus difficulty, which can be used when interpreting the own- versus other-race data from professionals.

Participants were recruited through emails sent to professional forensic facial working groups. These included the relative committees of the Organization of Scientific Area Committees (OSAC), the Facial Identification Scientific Working Group (FISWG), and the European Network of Forensic Science Institutes (ENFSI). In both remote and in-person sessions, the demographic questionnaire asked for information about the age, sex, race/ethnicity of the participant. It also asked whether the participant had taken any form of the CFMT before. The exact questions asked are listed in Appendix C.

2.1.2 Student Testing

A total of 86 undergraduate students from The University of Texas at Dallas (UTD) participated in this study. Data collection took place during the Spring 2019 semester. Participants were recruited through the School of Behavioral and Brain Sciences online sign-up system and were compensated with research exposure credits. Participants were required to be at least 18 years of age and have normal- or corrected-to-normal vision. The analysis included 48 Caucasian participants (35 female), ranging from age 18 to 37 (mean age 21.72), and 38 East Asian participants (27 female), ranging from age 18 to 36 (mean age 20.78). All aspects of the study were conducted in accordance with the UTD Institutional Review Board protocol.

Student participants completed the study in person in a single experimental session that included all three tests, followed by a demographic survey (via

[1]Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

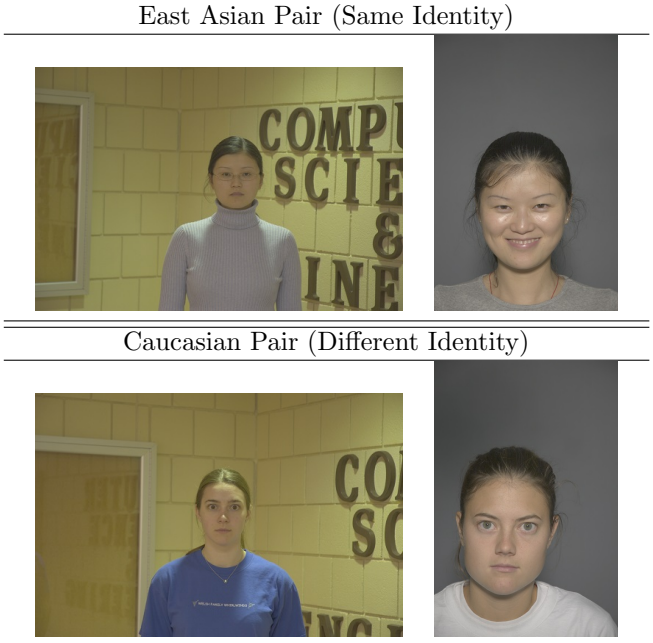


Figure 1: Example face pairs from the other-race identification experiment. The top pair is an example of an East Asian face pair; it is of the same identity. The bottom pair is an example of a Caucasian face pair; it is of different identities. Each pair contained an image with uncontrolled lighting (left images) and studio lighting (right images).

Qualtrics [41]). Test order was randomized across participants.

3 Other-Race Face Identification

3.1 Participants

In total, 118 participants participated in the test: (14 examiners, 18 reviewers, 48 Caucasian undergraduate students, and 38 East Asian undergraduate students). Data from 14 of 16 examiners were included in the analysis (one examiner did not complete the test; one examiner did not report Caucasian ancestry). None of the examiners reported Asian ancestry. Since all 18 reviewers reported some Caucasian ancestry and no Asian ancestry, we included all reviewers in the analysis. Caucasian and East Asian undergraduate students were recruited for the study.

3.2 Stimuli

Face images for this comparison were sourced from [15]. One image in the pair was taken under controlled illumination (*e.g.*, under studio lighting) and the other image was taken under uncontrolled illumination (*e.g.*, in a corridor). Example image pairs for the East Asian and Caucasian faces appear in Figure 1.

3.3 Procedure

Methods were adapted from [15]. Participants viewed each image pair for 30 seconds. Participants viewed four alternating blocks of 20 pairs of face images of East Asian and Caucasian individuals, for a total of 40 pairs of East Asian faces and 40 pairs of Caucasian faces. The order of stimuli in each block was fixed. Participants were asked to rate the face pairs on a 5-point scale. The scale offered the following response options: +2: Sure they are the same; +1: Think they are the same; 0: Do not know; −1: Think they are not the same; −2: Sure they are not the same. If the participant did not enter a response within 30 seconds, the image pair disappeared. The next image pair appeared when the participant provided a response.

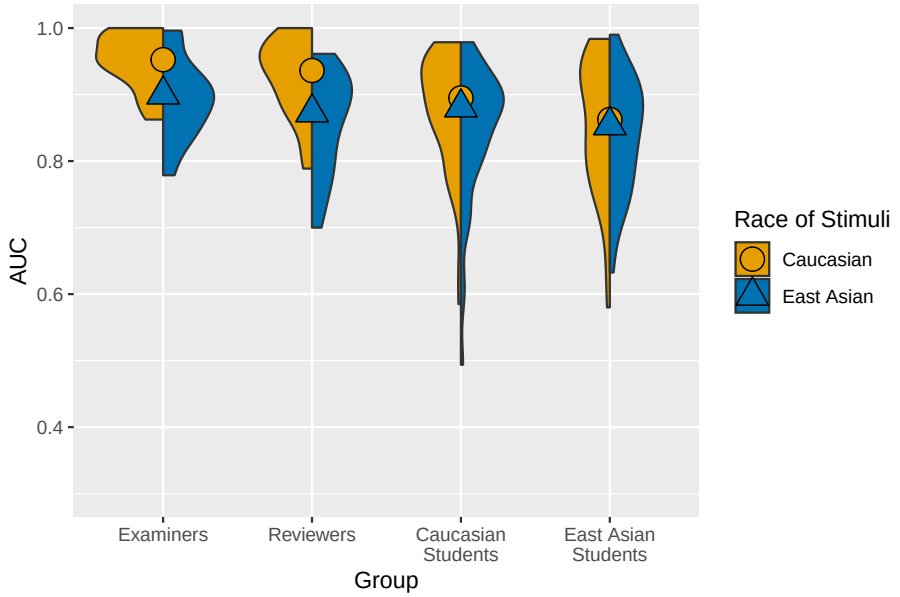
3.4 Results

Accuracy was measured separately for Caucasian and East Asian face pairs, using area under the receiver operating characteristic (ROC) curve (AUC) for each participant. Figure 2 shows the distributions of AUC for each participant group and stimulus race. The test was designed with the goal of applying a general linear model analysis (ANOVA) to assess results. The data, however, did not meet the basic pre-requisite conditions of normality and homogeneity of variance assumptions for parametric analyses. Therefore, we applied non-parametric Mann-Whitney tests to compare across participant groups on Caucasian and East Asian stimuli and Paired Wilcoxon comparisons for examining the effect of the stimulus race within each participant group. In all comparisons reported, p values have been Bonferroni-corrected[3] to account for multiple comparisons. Therefore, significance is still based on an α of 0.05.

We begin with the participant group comparisons for the Caucasian and East Asian face pairs (see Figure 2). For the Caucasian face pairs, three participant groups differed significantly. The first table in Figure 2 lists statistically significant comparisons and associated p values. Examiner performance was more accurate than both groups of students (East Asian, Caucasian). Reviewer performance was more accurate than the performance of East Asian students. For East Asian faces, performance did not differ across participant groups.

Next, we compared performance on the Caucasian and East Asian face pairs within each participant group. Both examiners and reviewers performed more

[3]For convenience and ease of interpretation, we multiplied each vector of p-values by n instead dividing α by n. Appendix A lists test statistics, medians, original p-values, and Bonferroni α-levels.



Groups		p-values	
		C Stimuli	EA Stimuli
Examiners	C Students	0.00735	ns.[2]
Examiners	EA Students	0.00342	ns.
Reviewers	EA Students	0.0459	ns.

Group	p-values
Examiners	0.000977
Reviewers	0.00336

Figure 2: Accuracy for Caucasian (C) and East Asian (EA) faces as a function of participant group in the other-race test. The distribution of AUCs for the Caucasian face pairs (orange) and East Asian (blue) are indicated, with medians shown using embedded shapes (circle/triangle). The top table shows comparisons between participant groups for Caucasian and East Asian face pairs with Mann-Whitney Bonferroni-corrected p-values. Examiners performed more accurately than students for Caucasian, but not East Asian, faces. Reviewers performed more accurately than students for Caucasian, but not East Asian, faces. The bottom table shows performance comparisons for Caucasian and East Asian faces within each participant group (paired Wilcoxon Bonferroni-corrected p-values). Each table only displays p-values that are significant at $\alpha = 0.05$. See Appendix A for all p-values and Bonferroni α-levels. Examiners and reviewers were more accurate on Caucasian face pairs, indicating an own-race advantage.

accurately on Caucasian face pairs than on East Asian face pairs. Neither student group's performance differed as a function of the face pair race.

In summary, examiners outperformed the Caucasian students on the Caucasian stimuli, replicating results found in literature [1, 3]. Although reviewers surpassed East Asian students identifying Caucasian face pairs, they were not more accurate than Caucasian students identifying East Asian face pairs. Both examiners and reviewers fared better with faces of their own race (Caucasian) than with faces of the other-race (East Asian). Because Caucasian students did not show this difference, we can conclude that examiners were more affected than students by the change from own-race to other-race face recognition–even if they were more accurate overall than the students.

4 Disguise

In this experiment, we compared performance of examiners, reviewers, and students (Caucasian and East Asian) on identification of non-disguised faces and two types of disguised faces (evasion and impersonation).

4.0.1 Participants

In total, the final analysis included 80 participants (14 examiners, 18 reviewers, 48 Caucasian students, and 38 East Asian students). Two examiners did not complete the test; all 14 examiners who completed the test are included in the analysis. Although we were not specifically focused on examining the variable of participant race for students, we retained both groups of participants and report on their results separately (see below).

4.0.2 Stimuli

The Façade dataset [37] includes two types of disguise: impersonation and evasion. With an impersonation disguise, one dataset subject aims to appear as another subject. With an evasion disguise, a subject attempts to appear differently from themselves in order not to be identified. Dataset subjects constructed their disguises themselves and were able to request items to aid their disguises from the researchers. Disguises were everyday wear and could not occlude the face (*e.g.*, no sunglasses); see [37] for more details.

Figure 3 shows examples of pair types from the Façade dataset. There are four types of face image pairs: same identity with no disguise, different identities with no disguise, evasion (*i.e.*, same identity with disguise), and impersonation (*i.e.*, different identities with disguise).

4.0.3 Procedure

The procedure we used was similar to that used in [37, 38], except that we used a response rating scale instead of the binary response used in the previous studies

(same identity, different identities). Specifically, we measured accuracy using the same 5-point scale, used in the other-race experiment (see Section 3.3).[4]

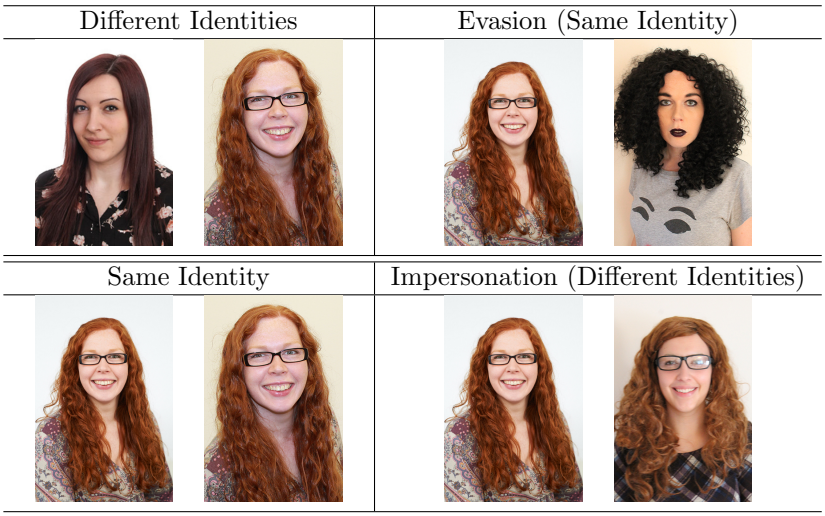


Figure 3: Example of images and pair types from the Façade dataset. In all pairs, the left image is the work profile photograph. The first column shows two examples of image pairs under the *non-disguised condition*: no disguises in any image pair. The top right shows an example of an image pairs in the *evasion condition*. The bottom right row shows an example of an image pair in the *impersonation condition*.

Each condition contained same- and different-identity image pairs. To compare accuracy on the dataset with previous studies on forensic facial professionals [1, 3], the *non-disguised condition* contained same- and different-identity pairs with no disguise. The evasion and impersonation conditions were used to test identification with disguise. The *evasion condition* contained evasion pairs (same-identity pairs composed of an undisguised identity and its evasion-disguised version) and different-identity pairs (undisguised faces from two different identities). The *impersonation condition* contained impersonation pairs (different-identity pairs composed of an undisguised identity and a person trying to resemble that identity) and same-identity pairs (undisguised face images of the same identity).

4.1 Results

Accuracy in each condition was assessed using AUC, computed for each participant. The graph in Figure 4 shows the distribution of accuracy in each

[4]Appendix B explores the binarized responses (*i.e.*, same or different) for comparibility to Noyes and Jenkins [37]. For comparability with White *et al.* [1] and Phillips *et al.* [3], we measured participant accuracy with AUC, area under the curve of the receiver operating characteristic (ROC).

group under all conditions. As we saw for the other-race experiment, the data did not meet pre-requisite conditions (normality and homogeneity of variance) for parametric analyses. Therefore, we applied non-parametric Mann-Whitney tests to compare the performance for participant groups on each condition (non-disguised, evasion, impersonation). We used Paired Wilcoxon comparisons for examining condition differences within each participant group. Again, p-values are corrected for multiple comparisons.

Beginning with the effects of the disguise manipulation within each group of participants, all participant groups were detrimentally affected by both impersonation and evasion disguises (compared to their performance on the non-disguised condition), and all groups performed lower on evasion disguises than on impersonation disguises (see the top table in Figure 4). The overall pattern of disguise effects is analogous to those reported previously [13].

Next, we compared across participant groups comparisons across each condition. The pattern of results here is more complex. Examiners were more accurate than East Asian students in all conditions (non-disguised, impersonation, and evasion). Examiners were more accurate than Caucasian students in the non-disguised and impersonation conditions, but not in the evasion condition. Examiners were more accurate than reviewers only in the evasion condition. Reviewers surpassed East Asian students, but only in the impersonation condition. The student groups performed comparably in all conditions.

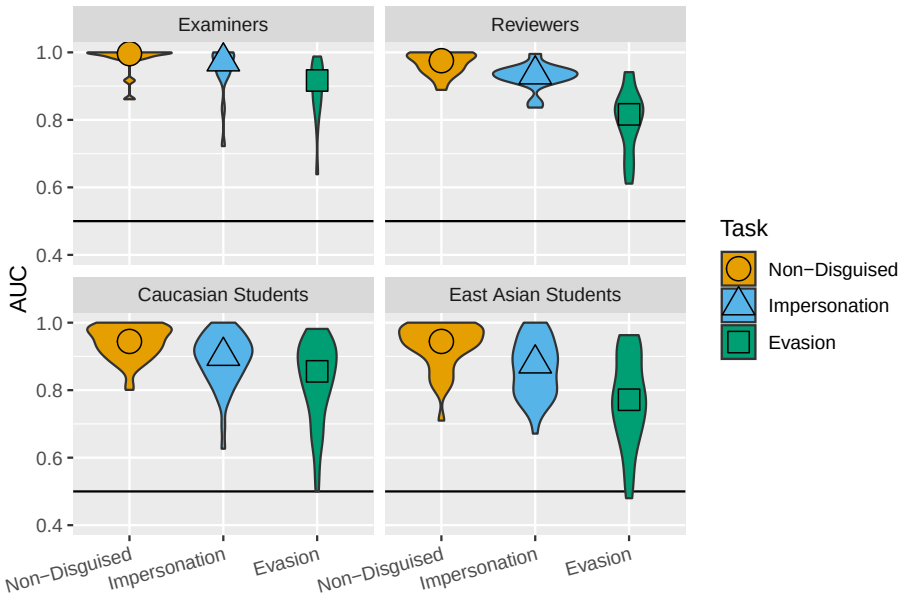
In summary, we show that the perceptual accuracy of forensic facial professionals is affected by the types of disguises in the same way as student's accuracy. All disguises adversely affected accuracy, and evasion was more challenging than impersonation. In comparisons between examiners and reviewers, and between reviewers and students, a more complex pattern of results emerged. In all cases, examiners performed more accurately than students.

5 Memory

In this test, we asked whether the skills of face examiners and reviewers extend to a face memory task. To address this question, examiners, reviewers, and students (Caucasian and East Asian) took the long form of the Cambridge Face Memory Test (CFMT+) [5]. The CFMT+ is able to differentiate between participants with superior face memory accuracy and those with typical memory [5].

5.0.1 Participants

A total of 78 participants completed the CFMT+ task: 13 examiners, 17 reviewers, 48 Caucasian students, and 38 East Asian students. The total number of examiners and reviewers are lower than the previous experiments due to the elimination of data from professionals who had taken a version of the CFMT previously (three examiners, one reviewer). The two student distributions were approximately normal and a one-way ANOVA found no difference between two



Conditions		p-values			
		Examiners	Reviewers	C Students	EA Students
N	I	0.02	0.00145	8.91×10^{-8}	1.84×10^{-6}
N	E	0.000366	2.29×10^{-5}	7.84×10^{-9}	3.56×10^{-7}
I	E	0.0157	4.58×10^{-5}	8.95×10^{-5}	3.28×10^{-5}

Groups		p-values		
		Non-Disguised	Impersonation	Evasion
Examiners	Reviewers	ns.	ns.	0.0247
Examiners	C Students	0.00765	0.0172	ns.
Examiners	EA Students	0.015	0.00768	0.0113
Reviewers	EA Students	ns.	0.0203	ns.

Figure 4: Group accuracy across non-disguised, impersonation, and evasion conditions. Median AUC for each group indicated with the smaller embedded shape. Note that chance performance is at $AUC = 0.50$ (indicated on graph with black line). The top table shows the effects of disguise on each participant. Impersonation (I) and evasion (E) disguise adversely affected all groups, relative to performance in the non-disguised control condition (N), and evasion proved more difficult than impersonation. The bottom table shows Bonferroni-corrected Mann-Whitney p-values comparing participant groups for each condition (non-disguised, impersonation, and evasion). Each table only displays p-values that are significant at $\alpha = 0.05$. See Appendix A for all p-values and Bonferroni α-levels.

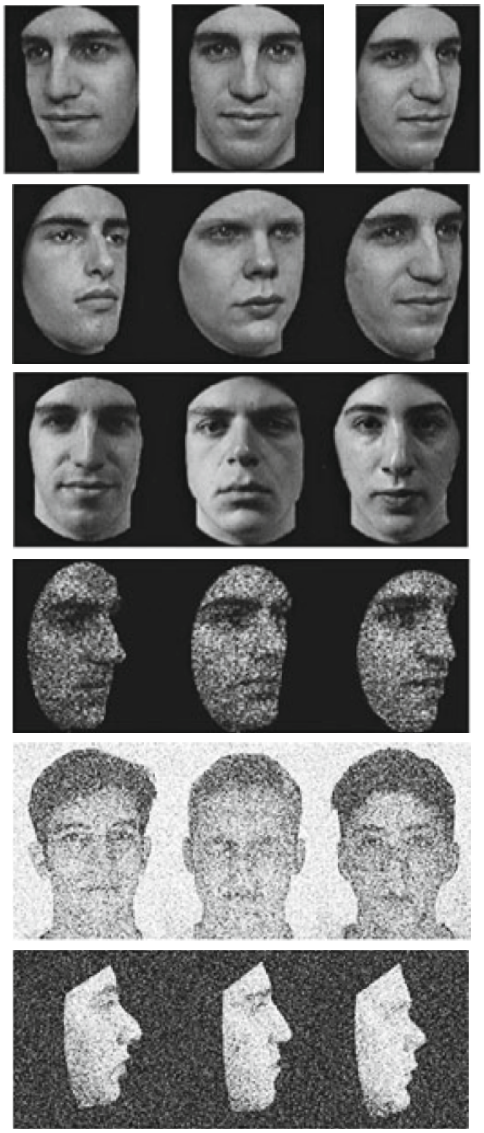


Figure 5: Example images from CFMT+. The first row shows the three angles participants see for 2 seconds to familiarize themselves with the identity. The remaining rows illustrate the images displayed in questions following memorization; the participant is asked to choose which of the three faces they were just asked to memorize. The last two rows indicate the harder trials present in the long form in order to detect high performers, *i.e.*, super-recognizers.

groups ($F(1, 84) = 0.1938$, $p = 0.6609$). Therefore, we combined the two student groups into a single participant group.

5.1 CFMT+ Test Protocol

The CFMT+ test was administered following its standard protocol [5]. In the first part, participants are shown an identity from three different angles; each angle is shown by itself for two seconds to familiarize themselves with the identity (see Figure 5, row 1). Once the participant has viewed all three angles, they are shown a row of three identities in one of the angles (see Figure 5, row 2). One identity is the one just viewed, and the other two are new identities. Participants are then asked to chose which identity they just viewed. For each of the six identities shown, the participant made three such decisions.

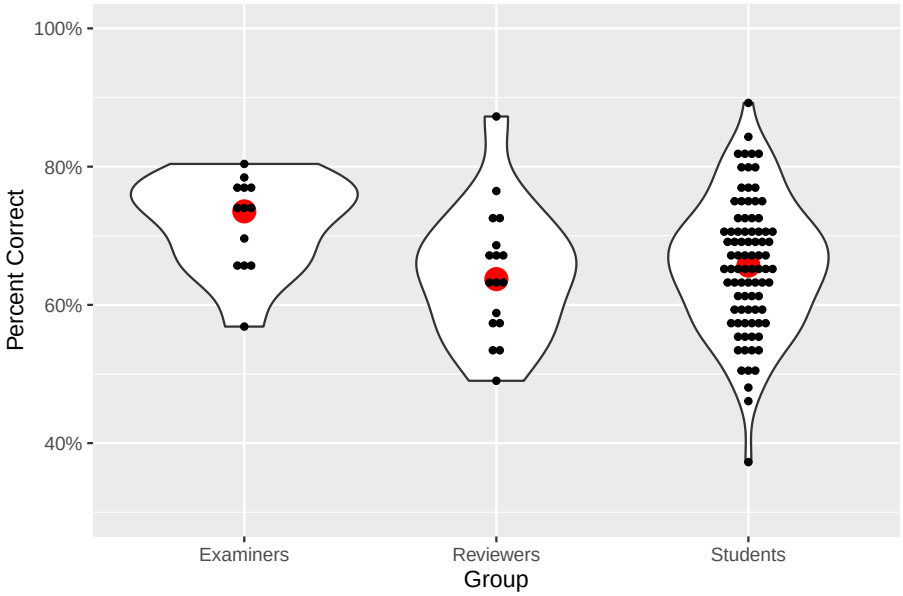
In the second part, participants are shown a 2×3 grid of 6 different identities from one angle, and they are given 20 seconds to memorize the faces. Afterwards, they are asked the same series of three alternative forced choice decisions and asked to choose which of the three identities present is an identity they have already seen. In the long form the CFMT, the trio of identities in the decision gets progressively more difficult, including adding visual static to the images to obscure features. See Figure 5 for an example; the last two rows (rows 5–6) are examples of more challenging trios present in CFMT+.

5.2 Results

Accuracy was measured as percent correct (PC). Figure 6 shows the distributions of accuracy for each group on the CFMT+. A one-way ANOVA between the groups ($F(2, 113) = 2.8291$, $p = 0.06326$) produces a p-value close to a cut-off of $\alpha = 0.05$. To investigate further and for consistency with the other two tests, the table in Figure 6 reports the Bonferroni Mann-Whitney p-values. For examiners, with both p-values slightly above significance (again, at $\alpha = 0.05$). Thus, this conclusion should be interpreted with caution. Reviewers and students performed comparably.

6 Discussion

Forensic facial examiners perform a critical role in face identifications and can present evidence in judicial proceedings. Previous studies of forensic professionals demonstrate their high levels of skill and accuracy at face identification [4, 1, 3]. Here, we expand the study of forensic facial professionals to include three challenging cases, with the goal of gaining insight into the nature of forensic professionals' face identification abilities. Before proceeding, we note that the administration of the other-race and disguise tests here did not allow forensic professionals access to the time and tools they typically use in casework. When given access to time and tools, previous work [1, 3] has demonstrated the perceptual accuracy to be a lower bound, *i.e.*, forensic professionals group



Groups		p-values
Examiners	Reviewers	0.0508
Examiners	Students	0.0512

Figure 6: Group accuracy on the CFMT+. In the graph, the x-axis indicates the group, and the y-axis is percent correct (PC). Each black dot represents an individual participant. The violin plot shows the density. The large red dots indicate the median PC for each group. The table shows the Mann-Whitney p-values comparing the groups. Examiners were marginally more accurate than reviewers and students on the CFMT+ test. Each table only displays p-values that are significant at $\alpha = 0.05$. See Appendix A for all p-values and Bonferroni α-levels.

accuracy does not lower with more time and tools. In what follows, we consider the results and implications of each experiment, in turn.

To begin, our results suggest that despite overall superiority in face identification, on this dataset, neither group of professionals was immune to the challenges of identifying other-race faces. Both examiners and reviewers performed less accurately on other-race faces than on own-race faces. Students, who were less accurate than the examiners overall, were nonetheless equally accurate for own- and other-race faces in this experiment. The equal performance of students for Caucasian and East Asian faces, combined with less accurate performance examiners and reviewers on East Asian faces, offers strong evidence that forensic facial professionals' superiority with own-race faces does not generalize completely to other-race faces. Although it is not possible to know for certain, the absence of an other-race effect for students in this study could be due to the high diversity of the local population in which the experiment was conducted[5].

For the case of disguise, the results replicated previous work with untrained participants [13], and expanded our knowledge of the limits of forensic facial professionals' skills. All groups of participants showed decreased accuracy for identifying faces under disguise, and all groups performed worse on evasion than impersonation disguise. Examiner performance surpassed the performance of both groups of students on impersonation disguise. Examiners were more accurate than East Asian students on evasion disguise, but they were not more accurate than the Caucasian students. Thus, we conclude that the forensic abilities of examiners generalize better to the case of impersonation, than to the case of evasion. Although the picture of results for reviewers is complex, combined with the findings for the CFMT+ and other-race tests, they are consistent with the idea that reviewers' performance is sometimes, but not always, on par with examiners.

The CFMT+ face memory test tracks a skill that is generally more similar to the use of face recognition in our daily lives. Faces must be remembered and later recalled to distinguish between strangers and the people we know. This ability is a critical life skill. Examiner performance on this task was marginally better than the performance of reviewers and students. Considering this finding, in the context of the superiority of examiners in several perceptually-based face identification tasks, suggests that basic face memory skills do not underlie examiners' superior performance. It is possible the marginal advantage we found here indicates that forensic facial professionals with generally good face recognition skills self-select into professional forensic facial examiner jobs. It is possible also, that a subset of the skills that examiners learn for perceptual face matching, apply in part to the memory task, possibly at the time of encoding the face memory. These questions are of interests for future work.

Finally, in this study, the matching tests are perceptual with viewing time limited to at most 30 seconds; examiners conduct their forensic comparisons with much more time. Their performance here can be considered a lower bound

[5]The University of Texas at Dallas has a highly diverse student population.

for their accuracy on casework based on previous studies [1, 3]. Conducting a forensic facial examiner closed-box test would elucidate the effect of stimuli race on accuracy under conditions similar to casework

References

[1] D. White, P. J. Phillips, C. A. Hahn, M. Hill, and A. J. O'Toole, "Perceptual expertise in forensic facial image comparison," *Proceedings of the Royal Society B*, vol. 282, 2015.

[2] A. Towler, D. White, and R. I. Kemp, "Evaluating the feature comparison strategy for forensic face identification," *Journal of Experimental Psychology: Applied*, vol. 23, no. 1, pp. 47–58, 2017.

[3] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, R. Chellappa, D. White, and A. J. O'Toole, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proceedings of the National Academy of Sciences*, vol. 115, no. 24, pp. 6171–6176, 2018.

[4] D. White, A. Towler, and R. I. Kemp, "Understanding Professional Expertise in Unfamiliar Face Matching," in *Forensic Face Matching: Research and Practice*, pp. 62–88, Oxford University Press, 01 2021.

[5] R. Russell, B. Duchaine, and K. Nakayama, "Super-recognizers: People with extraordinary face recognition ability," *Psychonomic Bulletin & Review*, vol. 16, pp. 252—257, Apr. 2009.

[6] E. Noyes, P. Phillips, and A. O'Toole, *What is a super-recogniser?*, pp. 173–201. Face Processing: Systems, Disorders and Cultural Differences, United States: Nova Science Publishers Inc, Sept. 2017.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[8] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9, IEEE, 2016.

[9] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[10] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 17–24, IEEE, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[12] R. S. Malpass and J. Kravitz, "Recognition for faces of own and other race," *Journal of Personality and Social Psychology*, vol. 13, no. 4, pp. 330–334, 1969.

[13] E. Noyes, *Face Recognition in Challenging Situations*. PhD thesis, University of York, 2016.

[14] A. M. Megreya, D. White, and A. M. Burton, "The other-race effect does not rely on memory: Evidence from a matching task," *Quarterly Journal of Experimental Psychology*, vol. 64, no. 8, pp. 1473–1483, 2011.

[15] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Applied Perception (TAP)*, vol. 8, no. 2, pp. 1–11, 2011.

[16] L. Feng, J. Liu, Z. Wang, J. Li, L. Li, L. Ge, J. Tian, and K. Lee, "The other face of the other-race effect: an fmri investigation of the other-race face categorization advantage," *Neuropsychologia*, vol. 49, no. 13, pp. 3739–3749, 2011.

[17] B. L. Hughes, N. P. Camp, J. Gomez, V. S. Natu, K. Grill-Spector, and J. L. Eberhardt, "Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14532–14537, 2019.

[18] V. Natu, D. Raboy, and A. J. O'Toole, "Neural correlates of own-and other-race face perception: Spatial and temporal response differences," *NeuroImage*, vol. 54, no. 3, pp. 2547–2555, 2011.

[19] G. Anzures, P. C. Quinn, O. Pascalis, A. M. Slater, J. W. Tanaka, and K. Lee, "Developmental origins of the other-race effect," *Current directions in psychological science*, vol. 22, no. 3, pp. 173–178, 2013.

[20] D. J. Kelly, S. Liu, K. Lee, P. C. Quinn, O. Pascalis, A. M. Slater, and L. Ge, "Development of the other-race effect during infancy: Evidence toward universality?," *Journal of experimental child psychology*, vol. 104, no. 1, pp. 105–114, 2009.

[21] E. McKone, S. Stokes, J. Liu, S. Cohan, C. Fiorentini, M. Pidcock, G. Yovel, M. Broughton, and M. Pelleg, "A robust method of measuring other-race and other-ethnicity effects: The cambridge face memory test format," *PLOS ONE*, vol. 7, pp. 1–6, Oct. 2012.

[22] A. J. O'toole, K. A. Deffenbacher, D. Valentin, and H. Abdi, "Structural aspects of face recognition and the other-race effect," *Memory & Cognition*, vol. 22, no. 2, pp. 208–224, 1994.

[23] J. E. Chance and A. G. Goldstein, "The other-race effect and eyewitness identification.," 1996.

[24] N. Furl, P. J. Phillips, and A. J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive science*, vol. 26, no. 6, pp. 797–815, 2002.

[25] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE transactions on biometrics, behavior, and identity science*, vol. 3, no. 1, pp. 101–111, 2020.

[26] K. Krishnapriya, K. Vangara, M. C. King, V. Albiero, and K. Bowyer, "Characterizing the variability in face recognition accuracy relative to race," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[27] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, 2020.

[28] H. El Khiyari and H. Wechsler, "Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning," *Journal of Biometrics and Biostatistics*, vol. 7, p. 323, 2016.

[29] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test part 3: Demographic effects," Dec. 2019.

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[32] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 118–126, 2015.

[33] S. Bate, R. Bennetts, N. Hasshim, E. Portch, E. Murray, E. Burns, and G. Dudfield, "The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 45, no. 3, pp. 363–377, 2019.

[34] D. J. Robertson, J. Black, B. Chamberlain, A. M. Megreya, and J. P. Davis, "Super-recognisers show an advantage for other race face identification," *Applied Cognitive Psychology*, vol. 34, no. 1, pp. 205–216, 2020.

[35] K. Patterson and A. Baddeley, "When face recognition fails.," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 3, no. 4, p. 406, 1977.

[36] G. Righi, J. J. Peissig, and M. J. Tarr, "Recognizing disguised faces," *Visual Cognition*, vol. 20, no. 2, pp. 143–169, 2012.

[37] E. Noyes and R. Jenkins, "Deliberate disguise in face identification," *Journal of Experimental Psychology: Applied*, vol. 25, pp. 280–290, Feb. 2019.

[38] E. Noyes, C. J. Parde, Y. I. Colón, M. Q. Hill, C. D. Castillo, R. Jenkins, and A. J. O'Toole, "Seeing through disguise: Getting to know you with a deep convolutional neural network," *Cognition*, vol. 211, 2021.

[39] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, *shiny: Web Application Framework for R*, 2019. R package version 1.3.2.

[40] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, pp. 195–203, 2019.

[41] Qualtrics, "Qualtrics." https://www.qualtrics.com/.

A Statistics

A.1 Other-Race Face Identification

Group 1	Group 2	N_1	N_2	Median 1	Median 2	U	p	α
Examiners	Reviewers	14	18	0.952	0.936	96.0	0.26200	0.00833
Examiners	C Students	14	48	0.952	0.895	143.5	0.00123	0.00833
Examiners	EA Students	14	38	0.952	0.862	98.5	0.00057	0.00833
Reviewers	C Students	18	48	0.936	0.895	262.5	0.01490	0.00833
Reviewers	EA Students	18	38	0.936	0.862	189.5	0.00765	0.00833
C Students	EA Students	48	38	0.895	0.862	800.0	0.33200	0.00833

Table 1: Mann-Whitney statistics on comparisons between groups on the Caucasian stimuli on the Other-Race Face Identification test. All p-values are unaltered, and the α is Bonferroni-corrected.

Group 1	Group 2	N_1	N_2	Median 1	Median 2	U	p	α
Examiners	Reviewers	14	18	0.900	0.873	100.0	0.3330	0.00833
Examiners	C Students	14	48	0.900	0.881	244.0	0.1230	0.00833
Examiners	EA Students	14	38	0.900	0.854	166.0	0.0401	0.00833
Reviewers	C Students	18	48	0.873	0.881	409.0	0.7460	0.00833
Reviewers	EA Students	18	38	0.873	0.854	299.0	0.4560	0.00833
C Students	EA Students	48	38	0.881	0.854	828.5	0.4700	0.00833

Table 2: Mann-Whitney statistics on comparisons between groups on the East Asian stimuli on the Other-Race Face Identification test.. All p-values are unaltered, and the α is Bonferroni-corrected.

Group	N	Median C Stim	Median EA Stim	W	p	α
Examiners	14	0.952	0.900	1.0	0.000244	0.0125
Reviewers	18	0.936	0.873	14.0	0.000839	0.0125
C Students	48	0.895	0.881	362.5	0.021000	0.0125
EA Students	38	0.862	0.854	292.5	0.261000	0.0125

Table 3: Paired Wilcoxon signed rank statistics on comparisons between stimuli sets for each group on the Other-Race Face Identification test. All p-values are unaltered, and the α is Bonferroni-corrected.

A.2 Façade

Group 1	Group 2	N_1	N_2	Median 1	Median 2	U	p	α
Examiners	Reviewers	14	18	0.996	0.975	84.5	0.11500	0.00833
Examiners	C Students	14	48	0.996	0.944	144.5	0.00127	0.00833
Examiners	EA Students	14	38	0.996	0.944	119.5	0.00250	0.00833
Reviewers	C Students	18	48	0.975	0.944	292.0	0.04440	0.00833
Reviewers	EA Students	18	38	0.975	0.944	227.5	0.04520	0.00833
C Students	EA Students	48	38	0.944	0.944	864.5	0.68300	0.00833

Table 4: Mann-Whitney statistics on comparisons between groups on the Non-Disguised condition on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

Group 1	Group 2	N_1	N_2	Median 1	Median 2	U	p	α
Examiners	Reviewers	14	18	0.973	0.934	69.5	0.03320	0.00833
Examiners	C Students	14	48	0.973	0.901	158.5	0.00287	0.00833
Examiners	EA Students	14	38	0.973	0.879	109.5	0.00128	0.00833
Reviewers	C Students	18	48	0.934	0.901	299.0	0.05640	0.00833
Reviewers	EA Students	18	38	0.934	0.879	174.5	0.00338	0.00833
C Students	EA Students	48	38	0.901	0.879	725.5	0.10600	0.00833

Table 5: Mann-Whitney statistics on comparisons between groups on the Impersonation condition on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

Group 1	Group 2	N_1	N_2	Median 1	Median 2	U	p	α
Examiners	Reviewers	14	18	0.917	0.816	50	0.00412	0.00833
Examiners	C Students	14	48	0.917	0.856	226	0.06520	0.00833
Examiners	EA Students	14	38	0.917	0.772	119	0.00189	0.00833
Reviewers	C Students	18	48	0.816	0.856	332	0.15200	0.00833
Reviewers	EA Students	18	38	0.816	0.772	297	0.43500	0.00833
C Students	EA Students	48	38	0.856	0.772	662	0.03000	0.00833

Table 6: Mann-Whitney statistics on comparisons between groups on the Evasion condition on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

Condition 1	Condition 2	N	Median 1	Median 2	W	p	α
Non-Disguised	Impersonation	14	0.996	0.973	41.0	6.67×10^{-3}	0.01667
Non-Disguised	Evasion	14	0.996	0.917	0	1.22×10^{-4}	0.01667
Impersonation	Evasion	14	0.973	0.917	10.0	5.25×10^{-3}	0.01667

Table 7: Paired Wilcoxon signed rank statistics on comparisons between conditions for Examiners on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

Condition 1	Condition 2	N	Median 1	Median 2	W	p	α
Non-Disguised	Impersonation	18	0.975	0.934	35.0	4.82×10^{-4}	0.01667
Non-Disguised	Evasion	18	0.975	0.816	0	7.63×10^{-6}	0.01667
Impersonation	Evasion	18	0.934	0.816	1.0	1.53×10^{-5}	0.01667

Table 8: Paired Wilcoxon signed rank statistics on comparisons between conditions for Reviewers on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

Condition 1	Condition 2	N	Median 1	Median 2	W	p	α
Non-Disguised	Impersonation	48	0.944	0.901	167.0	2.97×10^{-8}	0.01667
Non-Disguised	Evasion	48	0.944	0.856	7	2.61×10^{-9}	0.01667
Impersonation	Evasion	48	0.901	0.856	180.5	2.98×10^{-5}	0.01667

Table 9: Paired Wilcoxon signed rank statistics on comparisons between conditions for Caucasian Students on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

Condition 1	Condition 2	N	Median 1	Median 2	W	p	α
Non-Disguised	Impersonation	38	0.944	0.879	58.5	6.15×10^{-7}	0.01667
Non-Disguised	Evasion	38	0.944	0.772	38	1.19×10^{-7}	0.01667
Impersonation	Evasion	38	0.879	0.772	97.5	1.09×10^{-5}	0.01667

Table 10: Paired Wilcoxon signed rank statistics on comparisons between conditions for East Asian Students on the Façade test. All p-values are unaltered, and the α is Bonferroni-corrected.

A.3 CFMT

Group 1	Group 2	N_1	N_2	Median 1	Median 2	U	p	α
Examiners	Reviewers	13	17	0.735	0.637	53.0	0.0169	0.01667
Examiners	Students	13	86	0.735	0.657	328.5	0.0171	0.01667
Reviewers	Students	17	86	0.637	0.657	669.5	0.5880	0.01667

Table 11: Mann-Whitney statistics on comparisons between groups on the CFMT+. All p-values are unaltered, and the α is Bonferroni-corrected.

B Façade

The test created by Noyes and Jenkins [37] consisted of 156 pairs of face images with participants making binary decisions about each pair. Participants were not timed. Results were analyzed as percent correct. For our study, we showed participants a subset of 72 pairs and asked the participants to rate the similarity of the faces on a 5-point scale. Each pair was displayed for up to 30 seconds before disappearing. Once a response was entered, the participant moved to the next image pair.

The response scale for this study is different from Noyes and Jenkins [37] because AUC was used instead of percent correct. In order to compare our results to the analogous Experiment 1 in [37], we binarized the similarity scores (s).

In Equation 1 the scores are binarized with 1 and 2 being a declared match and -2, -1, and 0 being a declared non-match. After binarizing the scores, we looked at the percent correct for each group on each set, seen in Table 12.

$$bin(s) = \begin{cases} \text{match} & \text{if } s \in \{1, 2\} \\ \text{non-match} & \text{if } s \in \{-2, -1, 0\} \end{cases} \tag{1}$$

Table 12: Binarized group accuracy on Façade. (positive) ND stands for "no disguise."

Set	Examiners	Reviewers	Caucasian Students	[37] Students
ND Match	0.960	0.932	0.905	0.950
ND Non-Match	0.964	0.948	0.889	0.920
Evasion	0.623	0.512	0.662	0.600
Impersonation	0.893	0.818	0.766	0.820

In Equation 2 the scores are binarized with 0, 1, and 2 being a declared match and -2 and -1being a declared non-match. After binarizing the scores, we looked at the percent correct for each group on each set, seen in Table 13.

$$bin(s) = \begin{cases} \text{match} & \text{if } s \in \{0, 1, 2\} \\ \text{non-match} & \text{if } s \in \{-2, -1\} \end{cases} \tag{2}$$

Table 13: Binarized group accuracy on Façade. (non-negative) ND stands for "no disguise."

Set	Examiners	Reviewers	Caucasian Students	[37] Students
ND Match	0.976	0.938	0.922	0.950
ND Non-Match	0.917	0.926	0.858	0.920
Evasion	0.790	0.574	0.718	0.600
Impersonation	0.806	0.769	0.725	0.820

C Professional Background Questions

Examiners and reviewers were asked the following background questions. For those taking the tests on SurveyGizmo, the questions were asked over the phone after reviewing the consent form and before they took any tests. For those taking the tests on NIST laptops, the questions were taken on a Shiny (v1.3.2 [39]) application after completing all tests.

1. What is your sex?
   - ◯ Female
   - ◯ Male
2. What is your age?
   - ◯ 18-29
   - ◯ 30-39
   - ◯ 40-49
   - ◯ 50-59
   - ◯ 60-69
   - ◯ 70-79
   - ◯ 80+
3. Select one.
   - ◯ Hispanic or Latino
   - ◯ Not Hispanic or Latino
4. Please select the racial category or categories with which you most closely identify. Select one or more.
   - ☐ American Indian or Alaska Native
   - ☐ Asian
   - ☐ Black or African American
   - ☐ Native Hawaiian or Other Pacific Islander

☐ White

5. Have you ever taken the Cambridge Face Memory Test (CFMT)?

◯ Yes

◯ No