*Applied Cognitive Psychology*

WILEY

RESEARCH ARTICLE OPEN ACCESS

# Perceptual Expertise of Forensic Examiners and Reviewers on Tests of Cross-Race and Disguised Face Identification and Face Memory

Amy N. Yates[1] | Jacqueline G. Cavazos[2] | Géraldine Jeckeln[3] | Ying Hu[4,5] | Eilidh Noyes[6] | Carina A. Hahn[1] | Alice J. O'Toole[3] | P. Jonathon Phillips[1]

[1]National Institute of Standards and Technology, Gaithersburg, Maryland, USA | [2]University of California, Irvine, Irvine, California, USA | [3]The University of Texas at Dallas, Richardson, Texas, USA | [4]State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China | [5]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China | [6]School of Psychology, University of Leeds, Leeds, UK

**Correspondence:** P. Jonathon Phillips (jonathon.phillips@nist.gov)

**Received:** 4 December 2023 | **Revised:** 26 September 2024 | **Accepted:** 28 September 2024

**Funding:** This work was supported by National Institute of Standards and Technology, 70NANB21H109, 70NANB22H150, National Eye Institute, R01EY029692-04.

**Keywords:** disguised face identification | face matching | face memory | face recognition | other-race effect | other-race face identification

## ABSTRACT

Forensic facial professionals have been shown in previous studies to identify people from frontal face images more accurately than untrained participants when given 30 s per face pair. We tested whether this superiority holds in more challenging conditions. Two groups of forensic facial professionals (examiners, reviewers) and untrained participants were tested in three lab-based tasks: other-race face identification, disguised face identification, and face memory. For other-race face identification, on same-race faces, examiners were superior to controls; on different-race identification, examiners and controls performed comparably. Examiners were superior to controls for impersonation disguise, but not consistently superior for evasion disguise. Examiners' performance on the Cambridge Face Memory Test (CFMT+) was marginally better than reviewers and controls. We conclude that under laboratory-style conditions, professional examiners' identification superiority does not generalize completely to other-race and disguised faces. Future work should administer other-race and disguise face identification tests that allow forensic professionals to follow methods and procedures they typically use in casework.

## 1 | Introduction

Face identification judgments made by two groups of forensic facial professionals (examiners and reviewers) play an important role in law enforcement. *Facial examiners* perform detailed and careful comparisons of face images to determine whether the same person appears in two or more images (e.g., crime scene image and a mugshot). Due to their skill and training, the face identification judgments of examiners can be used as evidence in legal proceedings. *Facial reviewers* perform comparisons of face images typically under time constraints. The judgments of reviewers can be used to aid criminal investigations (e.g., by generating leads), but they cannot be used as evidence in legal proceedings. Despite the consequential nature of these roles, only in the last decade have there been studies comparing the accuracy of forensic facial examiners and reviewers to other untrained humans.

In recent years, multiple studies have reported controlled comparisons between novices and professionals on a variety of face identification tasks (Davis et al. 2016; Norell et al. 2015; Phillips et al. 2018; Robertson et al. 2016; Towler, White, and

Kemp 2017; Towler et al. 2019; White et al. 2015b). It is important to note that the criteria for inclusion as a "professional" and the difficulty of the stimulus set vary across these studies (cf., White, Towler, and Kemp 2021). Therefore, some caution is required in coming to conclusions based on individual studies. The comprehensive review and meta-analysis of comparisons between professionals and novices of White, Towler, and Kemp (2021) provides a valuable overview of this emerging literature. White, Towler, and Kemp (2021), show that forensic facial examiners are indeed more accurate than novices at face identification in the conditions tested (e.g., varying image quality, inverted images, images with body and clothing information). The superiority of facial reviewers over novices, however, is less consistently found—although this may in part be to differences in the way reviewers are defined across studies (White, Towler, and Kemp 2021). The performance of professionals (examiners, reviewers) also differs qualitatively from novices, with examiners/reviewers less disrupted by face inversion than novices, but more reliant on processing time (White et al. 2015b).[1] Specifically, these experts' face identification performance surpassed novice performance with 30 s. exposure times, but not with 2 s. exposure times (White et al. 2015b).

To date, studies comparing forensic examiners and reviewers to novices have concentrated on relatively standard tests of face identity matching. These studies have not included some challenging conditions that are encountered commonly in forensic casework. For example, evaluations of forensic examiners have primarily tested Caucasian examiners for identification of Caucasian faces. It is well known that face recognition is prone to error when untrained people are asked to recognize "other-race" faces—a phenomenon known as the cross-race effect (CRE) (Malpass and Kravitz 1969). The CRE has been replicated in dozens of studies with a variety of methodologies and paradigms (perceptual: Megreya, White, and Burton 2011; Phillips et al. 2011; neural: Feng et al. 2011; Hughes et al. 2019; Natu, Raboy, and O'Toole 2011; developmental: Anzures et al. 2013; Kelly et al. 2009; memory McKone et al. 2012; O'Toole et al. 1994; eyewitness memory tasks: Chance and Goldstein 1996). It is not known whether the superior performance of examiners and reviewers generalizes to faces of another race. The first goal of the present study was 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 cross-race test with equal numbers of Caucasian and East Asian face image pairs (Phillips et al. 2011).

Disguised faces also pose identification challenges for forensic casework. Disguise can have dramatic detrimental effects on face identification performance for untrained participants (Noyes 2016; Patterson and Baddeley 1977; Righi, Peissig, and Tarr 2012). In the most comprehensive study to date, Noyes and Jenkins created the *Façade* database of realistic facial disguises (Noyes and Jenkins 2019). 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). The results showed that unfamiliar observers were less accurate at identifying people with evasion and impersonation disguise. Evasion disguise proved especially difficult, with large decrements in identification performance. These results held regardless of whether observers were informed about the disguise manipulation. It is not known whether the skills of forensic examiners and reviewers generalize to the problem of facial disguise—although it is likely a condition encountered in forensic casework. The second goal of the present study was to compare professionals with untrained students on the task of face identification under disguise. To investigate this effect, we used the *Façade* database images.

The third goal of this study was to examine whether forensic facial experts' skills extend beyond perceptual identity matching to facial memory. Studies from the face processing literature point to the possibility that face perception skills and face memory skills may be dissociable. Results from Weigelt et al. (2014), for example, suggest that face perception develops in a domain-general fashion, along with the perception of other categories of objects, whereas face memory may develop and mature later along a domain-specific trajectory. However, the face perception and face memory skills of super-recognizers are related (Bobak, Hancock, and Bate 2016). Super-recognizers have been defined variously in the literature (Noyes, Phillips, and O'Toole 2017). In Bobak, Hancock, and Bate (2016), these individuals self-identified to the researchers as having excellent face recognition skills and were confirmed using a battery of face perception and recognition tests. Although super-recognizers and facial examiners both perform well on face identity matching tests, there is wide variability in the way super-recognizers are selected and pre-screened to qualify for this label (cf., White, Towler, and Kemp 2021). Moreover, forensic facial professionals are trained on perceptual face comparisons (Towler et al. 2019), whereas super-recognizers are not necessarily trained at all. It is unclear, therefore, whether forensic examiners, like super-recognizers, excel on face memory tasks. The third goal of this study was to determine whether the superior perceptual identity-matching skills of examiners generalize to a face memory task. To address this question, we tested examiners with the long form of the Cambridge Face Memory Test (CFMT+) (Russell, Duchaine, and Nakayama 2009)—a test widely used to separate people with superior face memory from those with from typical face memory. In this test, the identity comparison occurs 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. All three tests (cross-race face identification, disguised face identification, and face memory) were administered as laboratory style experiments. It is important to note that in forensic casework, examiners have access to tools and procedures that they can implement with ample time. In that context, the present study offers a first look at how professionals and untrained control participants compare on these challenging tasks. As such, it provides a lower bound estimate of the performance examiners might achieve in situ.

## 2 | Experiments

Participants completed three tests: (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. Therefore, the professionals are not subdivided by race. We begin with an overview of the participant groups and procedures. Then we proceed with a description of the three experiments.

## 3 | Participants and Test Administration

### 3.1 | Forensic Facial Professional Testing

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). 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. Self-reported data were used to determine eligibility for the study. 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 F. 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 complete.

Professional participants tested prior to May 2018 completed the task remotely. Researchers at NIST emailed task links to participants via Survey Gizmo.[2] 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 4 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 s. 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 Russell, Duchaine, and Nakayama (2009) 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 tests were followed by a demographic survey via Shiny v1.3.2 (Chang et al. 2019). The other-race test, disguise test, and CFMT+ were administered with PyschoPy v3.1.5 (Peirce et al. 2019).

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.

### 3.2 | Student Testing

A total of 86 undergraduate students from The University of Texas at Dallas 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 University of Texas at Dallas 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 Qualtrics (Qualtrics, n.d.). Test order was randomized across participants.

## 4 | Other-Race Face Comparisons

### 4.1 | 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. Because all 18 reviewers reported some Caucasian ancestry and no Asian ancestry, we included all reviewers in the analysis.

### 4.2 | Stimuli

Face images for this comparison were sourced from Phillips et al. (2011). 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.

### 4.3 | Procedure

Methods were adapted from Phillips et al. (2011). Participants viewed each image pair for 30 s. 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 s, the image pair disappeared. The next image pair appeared when the participant provided a response.

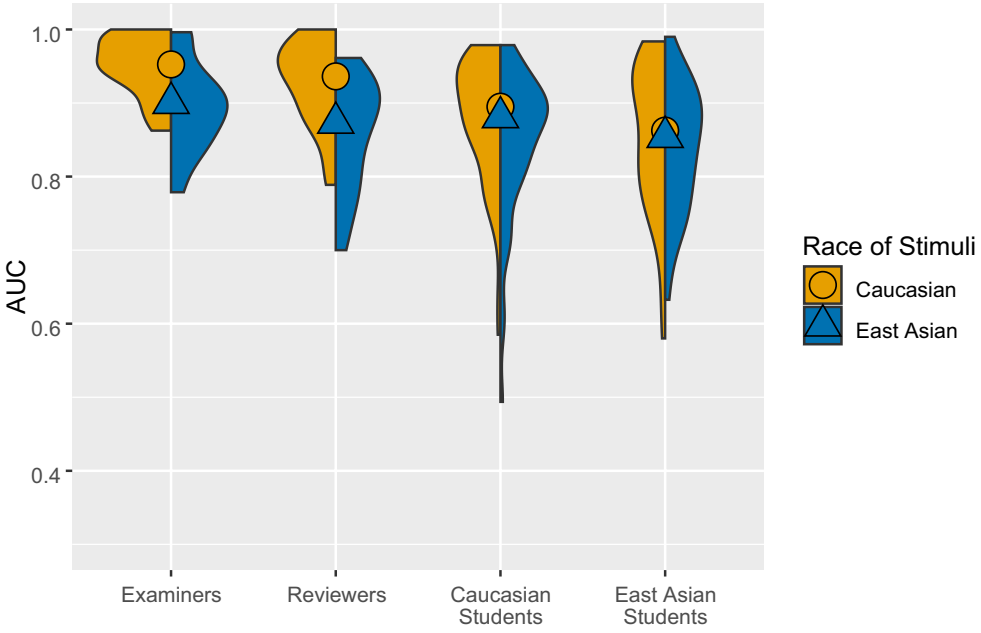
### 4.4 | Results

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 (i.e., 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.

We begin with the participant group comparisons for the Caucasian and East Asian face pairs, separately (see Figure 2). For the Caucasian face pairs, three participant groups differed significantly. The first table in Figure 2 lists the comparisons tested 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 as a function of the participant group.

Next, we compared performance on the Caucasian and East Asian face pairs within each participant group, the bottom table in Figure 2. Both examiners and reviewers performed


**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).




| | | *p*-values | |
|---|---|---|---|
| Group | | **C Stimuli** | **EA Stimuli** |
| **Examiners** | C Students | 0.007 | *ns.* |
| **Examiners** | EA Students | 0.003 | *ns.* |
| **Examiners** | Reviewers | *ns.* | *ns.* |
| **Reviewers** | C Students | *ns.* | *ns.* |
| **Reviewers** | EA Students | 0.046 | *ns.* |
| **C Students** | EA Students | *ns.* | *ns.* |

| Group | *p*-values |
|---|---|
| **Examiners** | < 0.001 |
| **Reviewers** | 0.003 |
| **C Students** | *ns.* |
| **EA Students** | *ns.* |

**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 East Asian 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). Examiners and reviewers were more accurate on Caucasian face pairs, indicating an own-race advantage. Tables display *p*-values that are significant at $\alpha = 0.05$. See Appendix A for all *p*-values and Bonferroni $\alpha$-levels.

more accurately on Caucasian face pairs than on East Asian face pairs. Neither student group's performance differed as a function of the race of the face pair. See Appendix A for statistics on comparisons between groups on the Caucasian stimuli (Table A1) and East Asian stimuli (Table A2) and between stimuli sets for each participant group (Table A3). See Appendix B (Tables B1 and B2) for results associated with binarized responses.

### 4.5 | Cross-Race Test Conclusions

In summary, examiners outperformed the Caucasian students on the Caucasian stimuli, replicating previous results (Phillips et al. 2018; White et al. 2015b). 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). Caucasian students did not show this difference, and so in the overall context of the results, we conclude that examiners were more affected than students by the change from own-race to other-race face recognition. The lack of a cross-race effect finding for students was unexpected. In the original test of students by Phillips et al. (2011), a cross-race effect was found (note: the participants of both tests were students from the University of Texas at Dallas). We consider possible explanations for this difference in the Discussion.

In summary, our results suggest that despite overall superiority in face identification in this test, neither group of professionals was immune to the challenges of identifying other-race faces.

## 4.6 | Disguised Face Comparisons

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).

### 4.6.1 | 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 participant race for students, we tested both the East Asian and Caucasian participants. For comparison to the other-race perceptual identity matching experiment, we did not combine the two groups.

### 4.6.2 | Stimuli

The *Façade* dataset (Noyes and Jenkins 2019) includes two types of disguise: impersonation and evasion. With impersonation disguise, a dataset subject aims to appear similar to a particular other subject. With evasion disguise, a dataset subject attempts to appear as "different" from themselves to evade identification. Dataset subjects constructed their disguises themselves and were able to request items to aid their disguises from the researchers. Disguises were everyday wear that did not occlude the face (e.g., no sunglasses); see Noyes and Jenkins (2019) for more details. Note that all face images in this experiment were of Caucasian individuals.

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). The pair types were assembled into three different conditions: non-disguised, evasion, and impersonation. Each condition contained same- and different-identity image pairs. The *non-disguised condition* (the left column in Figure 3) contained same- and different-identity pairs with no disguise. The evasion and impersonation conditions were used to test identification with disguise. The *evasion condition* (top row in Figure 3) 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* (bottom row in Figure 3) 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). See Appendix A for statistics on comparisons between groups on Non-Disguised (Table A4), Impersonation (Table A5), Evasion (Table A6) and comparisons between

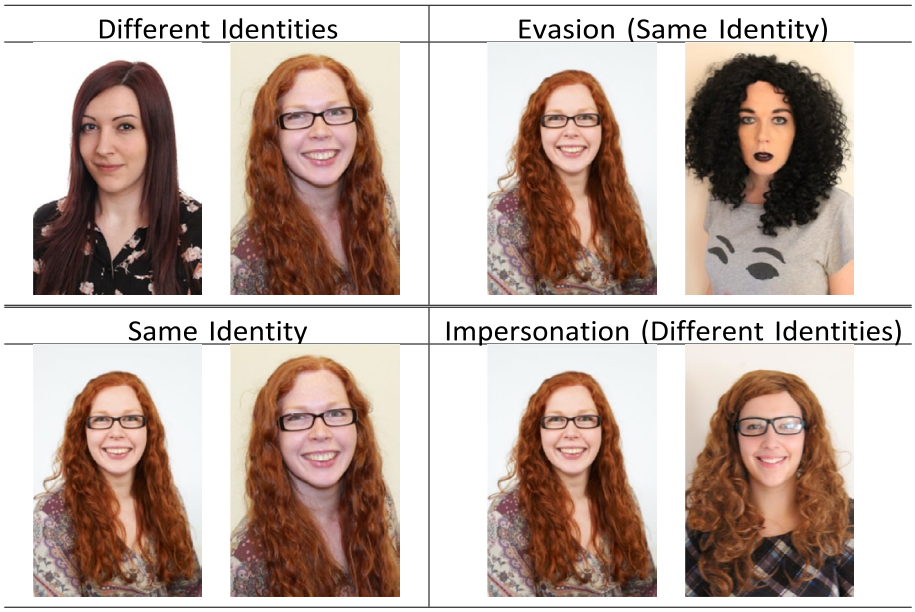



**FIGURE 3** | Example of images and pair types from the *Façade* dataset. In all pairs, the left image is the work profile photograph. The first column shows two examples of image pairs under the *non-disguised condition*: No disguises in any image pair. The top row shows an example of an image pairs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

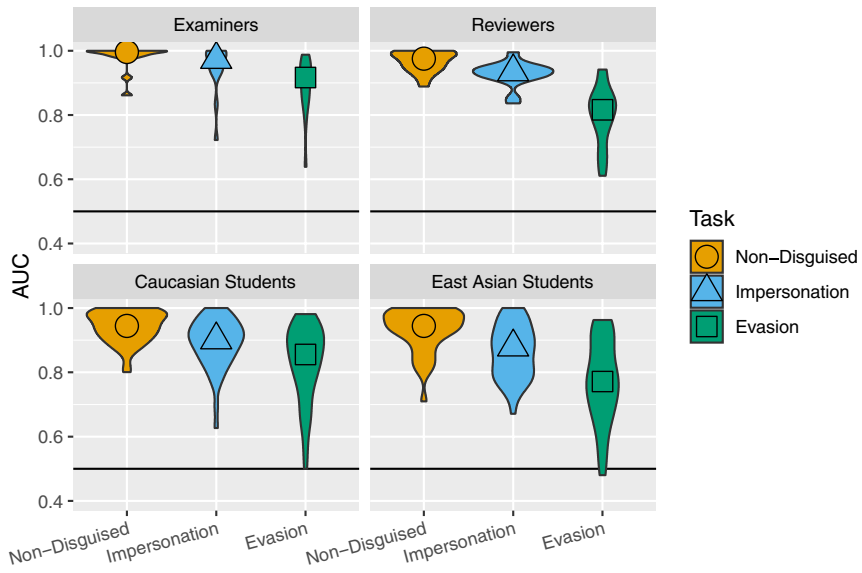
conditions for each participant group (Tables A7–A10). See Appendix C (Tables C1 and C2) for results associated with binarized responses.

#### 4.6.3 | Procedure

We used a procedure similar to that employed in Noyes and Jenkins (2019), with the exception that we used a response rating scale instead of a binary response (same identity, different identities). Specifically, we measured accuracy using the same 5-point scale, used in the other-race experiment (see the Other-Race Face Comparisons Procedure section).[4] Image pairs were presented in a fixed random order; they were not blocked by pair type, following the procedure of Noyes and Jenkins (2019). Noyes and Jenkins (2019) found no difference in performance regardless of whether participants were informed of the possibility of disguises. Here, participants were not informed that some images may contain a person wearing a disguise.

### 4.7 | 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 group under all conditions. The data did not meet pre-requisite conditions for parametric analyses (normality and homogeneity of variance). Therefore, we applied non-parametric Paired Wilcoxon comparisons for examining condition differences within each participant group. Again, *p*-values were corrected for multiple comparisons. All




| Conditions | | *p*-values Examiners | Reviewers | C Students | EA Students |
|---|---|---|---|---|---|
| Non-disguised | Impersonation | 0.02 | 0.001 | < 0.001 | < 0.001 |
| Non-disguised | Evasion | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| Impersonation | Evasion | 0.016 | < 0.001 | < 0.001 | < 0.001 |

| Groups | | *p*-values Non-Disguised | Impersonation | Evasion |
|---|---|---|---|---|
| Examiners | Reviewers | ns. | ns. | 0.025 |
| Examiners | C Students | 0.008 | 0.017 | ns. |
| Examiners | EA Students | 0.015 | 0.008 | 0.011 |
| Reviewers | C Students | *ns.* | *ns.* | *ns.* |
| Reviewers | EA Students | *ns.* | 0.020 | ns. |
| C Students | EA Students | *ns.* | *ns.* | *ns.* |

**FIGURE 4** | Accuracy across non-disguised, impersonation, and evasion conditions. Median AUC for each group indicated with the smaller embedded shape. Chance performance is at $AUC = 0.50$ (indicated on graph with black line). The top table shows Paired Wilcoxon tests to compare the effect of condition in each group. Impersonation and evasion disguise adversely affected all groups, relative to performance in the non-disguised condition: Evasion proved more difficult than impersonation. The bottom table show table shows Bonferroni-corrected Mann–Whitney *p*-values comparing participant groups for each condition (non-disguised, impersonation, and evasion). Examiners were more accurate than all students on non-disguised and impersonation disguise, but not evasion disguise. Tables display *p*-values significant at $\alpha = 0.05$. See Appendix A for Bonferroni $\alpha$-levels.

participant groups were detrimentally affected by both impersonation and evasion disguises (compared to their performance on the non-disguised condition) (see the top table in Figure 4). Additionally, all groups were less accurate on the evasion condition than on the impersonation condition. The disguise effects mirror those reported previously in Noyes and Jenkins (2019).

Next, we compared participant groups comparisons on each condition, using the Mann–Whitney statistic with *p*-values corrected for multiple comparisons. The pattern of results here is more complex. Examiners were more accurate than both groups of students in the non-disguised and impersonation conditions. Examiners were not more accurate than Caucasian students in the evasion condition but were more accurate than the East Asian students in this condition. Examiners were more accurate than reviewers in the evasion condition. Reviewers surpassed East Asian students, but only in the impersonation condition. Caucasian students and East Asian students performed comparably in all three conditions.

### 4.8 | Disguise Test Conclusions

These results replicate previous work with untrained participants (Noyes and Jenkins 2019) and expand our knowledge of the limits of forensic facial professionals' skills. In summary, we show that the perceptual accuracy of forensic facial professionals is affected by disguise in the same way as students' accuracy. All disguises adversely affected accuracy, with evasion more challenging than impersonation. The comparisons between examiners and reviewers, and between reviewers and students, produced a more complex pattern of results. The combined results indicate that the forensic abilities of examiners generalize better to impersonation than evasion disguise.

### 4.9 | Face Memory

In this test, we asked whether the skills of face examiners and reviewers extend beyond face matching, specifically 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+) (Russell, Duchaine, and Nakayama 2009). The CFMT+ can differentiate between participants with superior face memory accuracy and those with typical memory (Russell, Duchaine, and Nakayama 2009).

#### 4.9.1 | 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 groups ($F(1,84) = 0.1934$, $p = 0.661$), and so we combined the two student groups into a single participant group.

### 4.10 | CFMT+ Test Protocol

The CFMT+ test was administered following its standard protocol (Russell, Duchaine, and Nakayama 2009). In the first part, participants are shown an identity from three different angles; each angle is presented alone for 2 s to familiarize the participant with the identity (see Figure 5, row 1). Once the participant has viewed all three angles, a row of three identities is presented with all images displayed in one of the angles (see Figure 5, row 2). One of these images shows the identity just viewed, and the

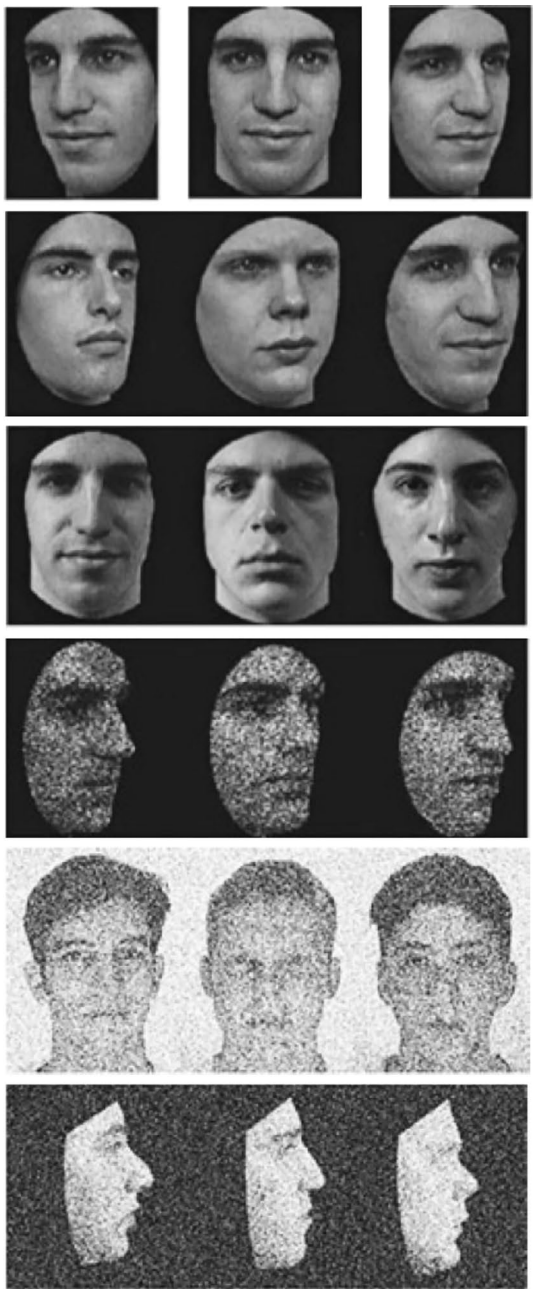


**FIGURE 5** | Example images from CFMT+. The first row shows the three pose angles participants see for 2 s to familiarize themselves with the identity. The remaining rows illustrate the images displayed in questions following memorization; the participant is asked to choose which of the three faces they were just asked to memorize. The last two rows show examples of the more difficult trials in the long form of the CFMT to detect high performers, that is, super-recognizers.

other two images show new identities. Participants are asked to choose which of the images shows the identity they just viewed. For each of the six identities shown, the participant made three such decisions.

In the second part, participants view a 2×3 grid of six different identities from one angle, and they are given 20 s to memorize the faces. Afterwards, they are asked the same series of three-alternative forced-choice decisions and are asked to select 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+.

### 4.11 | Results

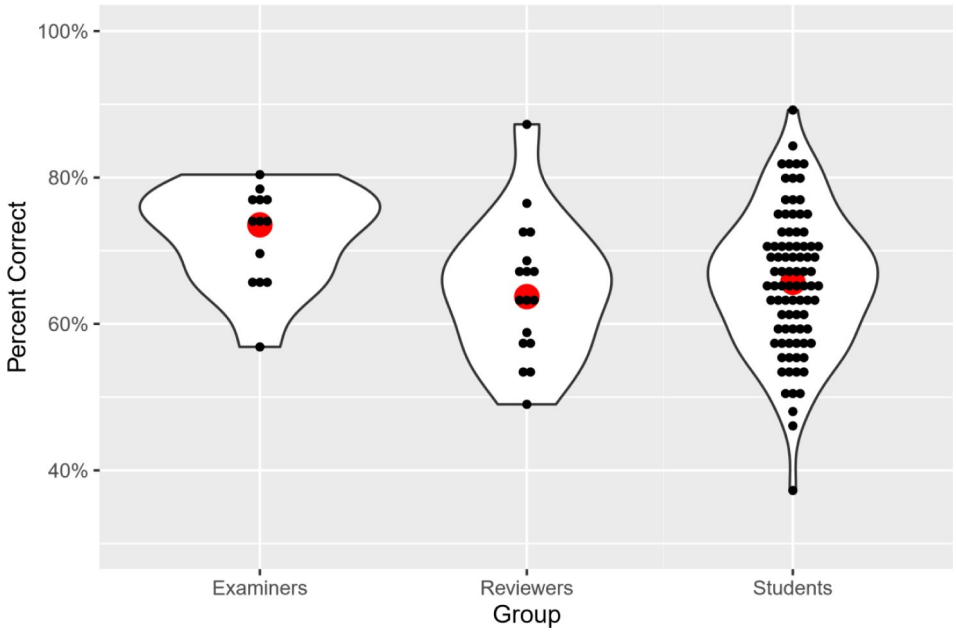
Accuracy was measured as percent correct for easier comparison across the tests. Figure 6 shows the distributions of accuracy for each group on the CFMT+; Appendix D draws Figure 6 with number correct for direct comparisons to other studies using the CFMT, which usually report results with number correct. A one-way ANOVA between the groups ($F(2,113) = 2.83$, $p = 0.06$) 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-corrected Mann–Whitney $p$-values. The comparisons between Examiners and Reviewers, and between Examiners and Students, both yielded $p$-values slightly above significance (again, at $\alpha = 0.05$) (see Figure 6 table). Thus, any conclusion should be interpreted with caution. Reviewers and students performed comparably. See Appendix A (Table A11) for statistics on comparisons between groups the CFMT+. See Appendix E (Table E1) for correlations between group performance on across all tests conducted in this study.

### 4.12 | Memory Test Conclusions

Examiner performance on this task was marginally better than the performance of reviewers and students. Figure 6 shows the high variability of performance on this task for all groups, but especially for the students. It is clear from Figure 6 that a number of the students performed more accurately than the best of the examiners. The finding of only a marginal difference suggests that examiners' face memory skills do not underpin their superior performance on perceptual face identification.

## 5 | Discussion

Forensic facial professionals perform an integral role in applied face identification scenarios. Previous studies of these professionals demonstrate their high levels of face identification




| Groups | | $p$-values |
|---|---|---|
| Examiners | Reviewers | 0.051 |
| Examiners | Students | 0.051 |
| Reviewers | Students | ns. |

**FIGURE 6** | Group accuracy on the CFMT+. In the graph, the $x$-axis indicates the group, and the $y$-axis is percent correct. 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. The table only displays $p$-values that are significant at $\alpha = 0.05$. See Appendix A for all $p$-values and Bonferroni $\alpha$-levels.

skill and accuracy but have employed tests that are limited in scope (Phillips et al. 2018; Towler et al. 2023; White, Towler, and Kemp 2021). In casework, examiners and reviewers must identify diverse faces from a variety of racial and ethnic backgrounds and must deal with attempts to evade identification by disguise. Face identification performance across racial diversity and deliberate attempts to evade identification by disguise are only a subset of the skills that should be investigated to obtain a complete picture of the performance of forensic face professionals.

Examining challenging face identification conditions can expand our understanding of the performance of high-performing groups and can move us closer to assuring the most accurate face identification possible in applied scenarios. Ultimately, the goal of comparing forensic professionals to untrained people is to determine the extent to which experts' identification decisions can be relied on in judicial proceedings and other scenarios where errors have serious consequences. The present study takes us part of the way to this goal, with the caveat that in casework, forensic professionals have access to their tools and procedures. They are also given ample time to apply these to the task at hand. The results of the present laboratory-style experiments show that there is a need to examine these challenging conditions using tests that enable a closer comparison between the performance of professionals in situ and the performance of novices. Our results offer the first evidence that we cannot assume that the performance of professionals will transfer to other-race face identification and to the identification of disguised individuals. These tasks tap skills that are commonly needed in forensic face identification casework. We consider each in turn.

For the question of cross-race identification, the present results replicate the superior performance of examiners on faces of their own race, while also showing that examiners identify other-race faces less accurately than own-race faces. This latter result is consistent with the well-documented cross-race effect for untrained observers (Malpass and Kravitz 1969). By contrast, the student control groups in this experiment did not show an other-race effect—an unexpected finding that does not replicate Phillips et al.'s (2011) earlier test of a student population from the same university on the same test. Notably, Caucasian and East Asian students also performed comparably on all conditions of the disguise experiment, which tested identification of Caucasian faces. Although we cannot offer a definite explanation of the difference between the present cross-race results for students and the results in Phillips et al. (2011), it is possible that in the decade-plus since that study, the University of Texas at Dallas campus and community may have become more diverse. This diversity may have attenuated the cross-race effect in students. A second possibility is that there is a generational/age difference between the (older) examiners and (college-age) students that may coincide with diversity difference in the community environments in which the two groups were raised. There is evidence to suggest that the roots of the own-race advantage for face recognition begin early in development (e.g., Kelly et al. 2009) and can be difficult to reverse—even with contact with more diverse individuals later in adulthood (cf., Ng and Lindsay 1994, although see Tanaka and Pierce 2009).

On the question of disguise, we show that examiners and reviewers are impaired in much the same way as untrained students. Impersonation and evasion disguise adversely affected all groups, relative to performance in the non-disguised control condition, and evasion proved more difficult than impersonation. This pattern of results mirrors that found in Noyes and Jenkins (2019) for participants unfamiliar with the individuals pictured. A similar pattern of results has also been reported for deep convolutional neural networks of face recognition when the training of the network was aimed at modeling unfamiliar face perception (Noyes et al. 2021; O'Toole and Castillo 2021). Notably, network performance improved markedly on both evasion and impersonation disguise with training aimed at mimicking familiar face perception. Specifically, this training simultaneously targeted identity differentiation and grouping across appearance variation within individual identities. This deep network training method was developed to bridge the performance gap between familiar and unfamiliar face identification (Jenkins et al. 2011). As computational models of face identification based on deep learning gain ground in challenging conditions, it will be important to consider their performance relative to the performance of experts and untrained participants (cf., Phillips et al. 2018).

Turning to the question of whether the perceptual superiority of examiners transfers to a memory test, the results of the present study are less clear. Examiners showed only a marginal advantage over reviewers and students on the CFMT. Arguably, face memory expertise is less important for examiners and reviewers than perceptual identity matching skills. From a more theoretical perspective, the lack of a solid memory advantage over novices could be due to a variety of factors. It is possible the small advantage we found here indicates that individuals with generally good face recognition skills self-select into professional face examiner jobs. It is possible also, that a subset of the skills that examiners learn for perceptual face matching, apply to the memory task, as well—possibly at the time of encoding. Regardless of the reasons underlying the weak effect, the present results suggests that examiners' face memory skills do not underpin their superior performance on perceptual face identification tasks.

From a broader perspective, computer-based face recognition has reached a level of accuracy that compares favorably with forensic examiners (Phillips et al. 2018; Towler et al. 2023). Moreover, computers now play a role in applied face recognition scenarios. Despite the level of performance achieved by machines, they too perform variably with faces of different races/ethnicities (Cavazos et al. 2021; El Khiyari and Wechsler 2016; Grother, Ngan, and Hanaoka 2019; Krishnapriya et al. 2020; Krishnapriya et al. 2019). And, as noted, machine performance declines with disguised faces (Noyes et al. 2021). Thus, machines are not necessarily a solution to the problem of identification in challenging conditions with serious consequences for errors. Overall, however, the impressive performance of machines should qualify them for inclusion in future evaluations of "experts."

In summary, we cannot assume that the skills of professional forensic examiners and reviewers generalize to cross-race face identification or to identification of disguised individuals

in real-life casework. Our laboratory-style experiments indicate that examiners and reviewers, like novices, are error-prone in the challenging conditions we investigated. Although laboratory-style experiments have limitations, much of what is currently known about the performance of experts compared to novices comes from laboratory-style tests (see Phillips et al. 2018). With this foundation in place, we can move forward to conduct evaluations that enable a closer evaluation of examiners in conditions similar to their working environment.

### Author Contributions

**Amy N. Yates:** conceptualization (equal), writing – original draft preparation (equal), writing – review and editing (equal), Investigation (lead), formal analysis (lead). **Jacqueline G. Cavazos:** conceptualization (equal), investigation (equal), formal analysis (equal). **Géraldine Jeckeln:** conceptualization (equal), writing – review and editing (equal), investigation (equal), formal analysis (equal). **Ying Hu:** conceptualization (equal), investigation (equal), formal analysis (equal). **Eilidh Noyes:** conceptualization (equal), investigation (equal), formal analysis (equal). **Carina A. Hahn:** conceptualization (equal), investigation (equal), formal analysis (equal). **Alice J. O'Toole:** conceptualization (lead), writing – original draft preparation (equal), writing – review and editing (equal), formal analysis (equal), supervision (equal). **P. Jonathon Phillips:** conceptualization (lead), writing – original draft preparation (equal), writing – review and editing (equal), formal analysis (equal), supervision (equal).


### Acknowledgments

The authors would like to thank Kaitie Karavai and Karen Marshall for their help in consenting forensic facial professionals. Research at the University of Texas at Dallas was funded by The National Institute of Standards and Technology, the National Eye Institute Grant R01EY029692–04 to AOT, Grant 70NANB21H109 & 70NANB22H150 to AOT.


### Ethics Statement

De-identified data from the student groups are available in the Open Science Framework site at the following link. In accordance with the IRB protocol, de-identified data from examiners and reviewers will be made available by e-mail request to the corresponding author via a data transfer agreement with NIST. Student data link: https://osf.io/a46mx/?view_only=18c91f1b87094322b2a1e3e6c330cdb1.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

De-identified data for forensic facial examiners and reviewers can be obtained by signing a data transfer agreement with the NIST. De-identified data for students will be available on Open Science Framework (OSF) upon acceptance. The images for the other-race face comparisons are available by license from the University of Notre Dame. The images for the Façade disguised face comparisons are available through Eilidh Noyes and Rob Jenkins. The images for the Cambridge Face Memory Test are available through Richard Russell.

### Endnotes

[1] Examiners and reviewers were combined into a single group in White et al. (2015).

[2] Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

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

[4] Appendix C explores the binarized responses (i.e., same or different) for comparability to Noyes and Jenkins (2019). For comparability with White et al. (2015) and Phillips et al. (2018), we measured participant accuracy with AUC, area under the curve of the receiver operating characteristic (ROC).

### References


Anzures, G., P. C. Quinn, O. Pascalis, A. M. Slater, J. W. Tanaka, and K. Lee. 2013. "Developmental Origins of the Other-Race Effect." *Current Directions in Psychological Science* 22, no. 3: 173–178.

Bobak, A. K., P. J. Hancock, and S. Bate. 2016. "Super-Recognisers in Action: Evidence From Face-Matching and Face Memory Tasks." *Applied Cognitive Psychology* 30, no. 1: 81–91.

Cavazos, J. G., P. J. Phillips, C. D. Castillo, and A. J. O'Toole. 2021. "Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias?" *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, no. 1: 101–111.

Chance, J. E., and A. G. Goldstein. 1996. "The other-race effect and eyewitness identification." In *Psychological issues in eyewitness identification*, edited by S. L. Sporer, R. S. Malpass and G. Koehnken, 153–176. New Jersey: Lawrence Erlbaum Associates, Inc.

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson. 2019. "Shiny: Web application framework for r [Computer software manual]." https://CRAN.R-project.org/package=shiny (R package version 1.3.2).

Davis, J. P., K. Lander, R. Evans, and A. Jansari. 2016. "Investigating Predictors of Superior Face Recognition Ability in Police Super-Recognisers." *Applied Cognitive Psychology* 30, no. 6: 827–840.

El Khiyari, H., and H. Wechsler. 2016. "Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning." *Journal of Biometrics and Biostatistics* 7: 323.

Feng, L., J. Liu, Z. Wang, et al. 2011. "The Other Face of the Other-Race Effect: An Fmri Investigation of the Other-Race Face Categorization Advantage." *Neuropsychologia* 49, no. 13: 3739–3749.

Grother, P., M. Ngan, and K. Hanaoka. 2019. "Face Recognition Vendor Test Part 3: Demographic Effects." In *NIST Interagency/Internal Report (NISTIR)*. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280.

Hughes, B. L., N. P. Camp, J. Gomez, V. S. Natu, K. Grill-Spector, and J. L. Eberhardt. 2019. "Neural Adaptation to Faces Reveals Racial Outgroup Homogeneity Effects in Early Perception." *Proceedings of the National Academy of Sciences* 116, no. 29: 14532–14537.

Jenkins, R., D. White, X. Van Montfort, and A. M. Burton. 2011. "Variability in Photos of the Same Face." *Cognition* 121, no. 3: 313–323.

Kelly, D. J., S. Liu, K. Lee, et al. 2009. "Development of the Other-Race Effect During Infancy: Evidence Toward Universality?" *Journal of Experimental Child Psychology* 104, no. 1: 105–114.

Krishnapriya, K., K. Vangara, M. C. King, V. Albiero, and K. Bowyer. 2019. "Characterizing the Variability in Face Recognition Accuracy Relative to Race." In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Krishnapriya, K., V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer. 2020. "Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone." *IEEE Transactions on Technology and Society* 1:

8–20. Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone.

Malpass, R. S., and J. Kravitz. 1969. "Recognition for Faces of Own and Other Race." *Journal of Personality and Social Psychology* 13, no. 4: 330–334.

McKone, E., S. Stokes, J. Liu, et al. 2012. "A Robust Method of Measuring Other-Race and Other-Ethnicity Effects: The Cambridge Face Memory Test Format." *PLoS One* 7, no. 10: 1–6.

Megreya, A. M., D. White, and A. M. Burton. 2011. "The Other-Race Effect Does Not Rely on Memory: Evidence From a Matching Task." *Quarterly Journal of Experimental Psychology* 64, no. 8: 1473–1483.

Natu, V., D. Raboy, and A. J. O'Toole. 2011. "Neural Correlates of Own-and Other-Race Face Perception: Spatial and Temporal Response Differences." *NeuroImage* 54, no. 3: 2547–2555.

Ng, W. J., and R. C. Lindsay. 1994. "Cross-Race Facial Recognition: Failure of the Contact Hypothesis." *Journal of Cross-Cultural Psychology* 25, no. 2: 217–232.

Norell, K., K. B. Läthén, P. Bergström, A. Rice, V. Natu, and A. O'Toole. 2015. "The Effect of Image Quality and Forensic Expertise in Facial Image Comparisons." *Journal of Forensic Sciences* 60, no. 2: 331–340.

Noyes, E. 2016. "Face Recognition in Challenging Situations." Unpublished Doctoral dissertation, University of York.

Noyes, E., and R. Jenkins. 2019. "Deliberate disguise in face identification." *Journal of Experimental Psychology: Applied* 25: 280–290. https://doi.org/10.1037/xap0000213.

Noyes, E., P. Phillips, and A. O'Toole. 2017. "What Is a Super-Recogniser?" In *Face Processing: Systems, Disorders and Cultural Differences*, edited by M. Bindemann and A. Megreya, 173–201. United States: Nova Science Publishers Inc.

Noyes, E., C. J. Parde, Y. I. Colón, et al. 2021. "Seeing Through Disguise: Getting to Know You With a Deep Convolutional Neural Network." *Cognition* 211: 104611.

O'Toole, A. J., and C. D. Castillo. 2021. "Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning." *Annual Review of Vision Science* 7: 543–570.

O'Toole, A. J., K. A. Deffenbacher, D. Valentin, and H. Abdi. 1994. "Structural Aspects of Face Recognition and the Other-Race Effect." *Memory & Cognition* 22, no. 2: 208–224.

Patterson, K., and A. Baddeley. 1977. "When Face Recognition Fails." *Journal of Experimental Psychology: Human Learning and Memory* 3, no. 4: 406–417.

Peirce, J., J. R. Gray, S. Simpson, et al. 2019. "Psychopy2: Experiments in Behavior Made Easy." *Behavior Research Methods* 51: 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Phillips, P. J., F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole. 2011. "An Other-Race Effect for Face Recognition Algorithms." *ACM Transactions on Applied Perception (TAP)* 8, no. 2: 1–11.

Phillips, P. J., A. N. Yates, Y. Hu, et al. 2018. "Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms." *Proceedings of the National Academy of Sciences* 115, no. 24: 6171–6176. https://www.pnas.org/content/115/24/6171. https://doi.org/10.1073/pnas.1721355115.

Qualtrics. n.d. "Qualtrics." https://www.qualtrics.com/. Provo, Utah, USA.

Righi, G., J. J. Peissig, and M. J. Tarr. 2012. "Recognizing Disguised Faces." *Visual Cognition* 20, no. 2: 143–169.

Robertson, D. J., E. Noyes, A. J. Dowsett, R. Jenkins, and A. M. Burton. 2016. "Face Recognition by Metropolitan Police Super-Recognisers." *PLoS One* 11, no. 2: e0150036.

Russell, R., B. Duchaine, and K. Nakayama. 2009. "Super-Recognizers: People With Extraordinary Face Recognition Ability." *Psychonomic Bulletin & Review* 16, no. 2: 252–257. https://doi.org/10.3758/PBR.16.2.252.

Tanaka, J. W., and L. J. Pierce. 2009. "The Neural Plasticity of Other-Race Face Recognition." *Cognitive, Affective, & Behavioral Neuroscience* 9, no. 1: 122–131.

Towler, A., D. White, and R. I. Kemp. 2017. "Evaluating the Feature Comparison Strategy for Forensic Face Identification." *Journal of Experimental Psychology: Applied* 23, no. 1: 47–58. https://doi.org/10.1037/xap0000108.

Towler, A., R. I. Kemp, A. M. Burton, et al. 2019. "Do Professional Facial Image Comparison Training Courses Work?" *PLoS One* 14, no. 2: e0211037.

Towler, A., J. D. Dunn, S. Castro Martínez, et al. 2023. "Diverse Types of Expertise in Facial Recognition." *Scientific Reports* 13: 11396. https://doi.org/10.1038/s41598-023-28632-x.

Weigelt, S., K. Koldewyn, D. D. Dilks, B. Balas, E. McKone, and N. Kanwisher. 2014. "Domain-Specific Development of Face Memory but Not Face Perception." *Developmental Science* 17, no. 1: 47–58.

White, D., P. J. Phillips, C. A. Hahn, M. Hill, and A. J. O'Toole. 2015b. "Perceptual Expertise in Forensic Facial Image Comparison." *Proceedings of the Royal Society B* 282: 20151292. https://doi.org/10.1098/rspb.2015.1292.

White, D., A. Towler, and R. I. Kemp. 2021. "Understanding Professional Expertise in Unfamiliar Face Matching." In *Forensic Face Matching: Research and Practice*, 62–88. New York, NY: Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0004.

## Appendix A

### A.1 | Statistics

**TABLE A1** | 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 $\alpha$ is Bonferroni-corrected.

| Group 1 | Group 2 | $N_1$ | $N_2$ | Median 1 | Median 2 | *U* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Examiners | Reviewers | 14 | 18 | 0.952 | 0.936 | 96.0 | 0.262 | 0.0083 |
| Examiners | C Students | 14 | 48 | 0.952 | 0.895 | 144.0 | 0.001 | 0.0083 |
| Examiners | EA Students | 14 | 38 | 0.952 | 0.862 | 98.5 | <0.001 | 0.0083 |
| Reviewers | C Students | 18 | 48 | 0.936 | 0.895 | 262.0 | 0.015 | 0.0083 |
| Reviewers | EA Students | 18 | 38 | 0.936 | 0.862 | 190.0 | 0.0077 | 0.0083 |
| C Students | EA Students | 48 | 38 | 0.895 | 0.862 | 800.0 | 0.332 | 0.0083 |

**TABLE A2** | 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 $\alpha$ is Bonferroni-corrected.

| Group 1 | Group 2 | $N_1$ | $N_2$ | Median 1 | Median 2 | *U* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Examiners | Reviewers | 14 | 18 | 0.900 | 0.873 | 100.0 | 0.333 | 0.0083 |
| Examiners | C Students | 14 | 48 | 0.900 | 0.881 | 244.0 | 0.123 | 0.0083 |
| Examiners | EA Students | 14 | 38 | 0.900 | 0.854 | 166.0 | 0.040 | 0.0083 |
| Reviewers | C Students | 18 | 48 | 0.873 | 0.881 | 409.0 | 0.746 | 0.0083 |
| Reviewers | EA Students | 18 | 38 | 0.873 | 0.854 | 299.0 | 0.456 | 0.0083 |
| C Students | EA Students | 48 | 38 | 0.881 | 0.854 | 828.0 | 0.470 | 0.0083 |

**TABLE A3** | 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 $\alpha$ is Bonferroni-corrected.

| Group | *N* | Median C Stim | Median EA Stim | *W* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|
| Examiners | 14 | 0.952 | 0.900 | 1.0 | <0.001 | 0.0125 |
| Reviewers | 18 | 0.936 | 0.873 | 14.0 | <0.001 | 0.0125 |
| C Students | 48 | 0.895 | 0.881 | 362.0 | 0.021 | 0.0125 |
| EA Students | 38 | 0.862 | 0.854 | 292.0 | 0.261 | 0.0125 |

**TABLE A4** | Mann–Whitney statistics on comparisons between groups on the Non-Disguised condition on the Façade test. All *p*-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Group 1 | Group 2 | $N_1$ | $N_2$ | Median 1 | Median 2 | *U* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Examiners | Reviewers | 14 | 18 | 0.996 | 0.975 | 84.5 | 0.115 | 0.0083 |
| Examiners | C Students | 14 | 48 | 0.996 | 0.944 | 144.0 | 0.001 | 0.0083 |
| Examiners | EA Students | 14 | 38 | 0.996 | 0.944 | 120.0 | 0.003 | 0.0083 |
| Reviewers | C Students | 18 | 48 | 0.975 | 0.944 | 292.0 | 0.044 | 0.0083 |
| Reviewers | EA Students | 18 | 38 | 0.975 | 0.944 | 228.0 | 0.045 | 0.0083 |
| C Students | EA Students | 48 | 38 | 0.944 | 0.944 | 864.0 | 0.683 | 0.0083 |

**TABLE A5** | Mann–Whitney statistics on comparisons between groups on the Impersonation condition on the Façade test. All *p*-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Group 1 | Group 2 | $N_1$ | $N_2$ | Median 1 | Median 2 | *U* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Examiners | Reviewers | 14 | 18 | 0.973 | 0.934 | 69.5 | 0.033 | 0.0083 |
| Examiners | C Students | 14 | 48 | 0.973 | 0.901 | 158.0 | 0.003 | 0.0083 |
| Examiners | EA Students | 14 | 38 | 0.973 | 0.879 | 110.0 | 0.001 | 0.0083 |
| Reviewers | C Students | 18 | 48 | 0.934 | 0.901 | 299.0 | 0.056 | 0.0083 |
| Reviewers | EA Students | 18 | 38 | 0.934 | 0.879 | 174.5 | 0.003 | 0.0083 |
| C Students | EA Students | 48 | 38 | 0.901 | 0.879 | 726.0 | 0.106 | 0.0083 |

**TABLE A6** | Mann–Whitney statistics on comparisons between groups on the Evasion condition on the Façade test. All *p*-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Group 1 | Group 2 | $N_1$ | $N_2$ | Median 1 | Median 2 | *U* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Examiners | Reviewers | 14 | 18 | 0.917 | 0.816 | 50.0 | 0.004 | 0.0083 |
| Examiners | C Students | 14 | 48 | 0.917 | 0.856 | 226.0 | 0.065 | 0.0083 |
| Examiners | EA Students | 14 | 38 | 0.917 | 0.772 | 119.0 | 0.002 | 0.0083 |
| Reviewers | C Students | 18 | 48 | 0.816 | 0.856 | 332.0 | 0.152 | 0.0083 |
| Reviewers | EA Students | 18 | 38 | 0.816 | 0.772 | 297.0 | 0.435 | 0.0083 |
| C Students | EA Students | 48 | 38 | 0.856 | 0.772 | 662.0 | 0.030 | 0.0083 |

**TABLE A7** | Paired Wilcoxon signed rank statistics on comparisons between conditions for Examiners on the Façade test. All *p*-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Condition 1 | Condition 2 | *N* | Median 1 | Median 2 | *W* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|
| Non-disguised | Impersonation | 14 | 0.996 | 0.973 | 41.0 | 0.007 | 0.0167 |
| Non-disguised | Evasion | 14 | 0.996 | 0.917 | 0 | $<0.001$ | 0.0167 |
| Impersonation | Evasion | 14 | 0.973 | 0.917 | 10.0 | 0.005 | 0.0167 |

**TABLE A8** | Paired Wilcoxon signed rank statistics on comparisons between conditions for Reviewers on the Façade test. All *p*-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Condition 1 | Condition 2 | *N* | Median 1 | Median 2 | *W* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|
| Non-disguised | Impersonation | 18 | 0.975 | 0.934 | 35.0 | $<0.001$ | 0.0167 |
| Non-disguised | Evasion | 18 | 0.975 | 0.816 | 0 | $<0.001$ | 0.0167 |
| Impersonation | Evasion | 18 | 0.934 | 0.816 | 1.0 | $<0.001$ | 0.0167 |

**TABLE A9** | Paired Wilcoxon signed rank statistics on comparisons between conditions for Caucasian Students on the Façade test. All *p*-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Condition 1 | Condition 2 | *N* | Median 1 | Median 2 | *W* | *p* | $\alpha$ |
|---|---|---|---|---|---|---|---|
| Non-disguised | Impersonation | 48 | 0.944 | 0.901 | 167.0 | $<0.001$ | 0.0167 |
| Non-disguised | Evasion | 48 | 0.944 | 0.856 | 7.0 | $<0.001$ | 0.0167 |
| Impersonation | Evasion | 48 | 0.901 | 0.856 | 180.0 | $<0.001$ | 0.0167 |

**TABLE A10** | Paired Wilcoxon signed rank statistics on comparisons between conditions for East Asian Students on the Façade test. All $p$-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Condition 1 | Condition 2 | $N$ | Median 1 | Median 2 | $W$ | $p$ | $\alpha$ |
|---|---|---|---|---|---|---|---|
| Non-disguised | Impersonation | 38 | 0.944 | 0.879 | 58.5 | $<0.001$ | 0.0167 |
| Non-disguised | Evasion | 38 | 0.944 | 0.772 | 38 | $<0.001$ | 0.0167 |
| Impersonation | Evasion | 38 | 0.879 | 0.772 | 97.5 | $<0.001$ | 0.0167 |

**TABLE A11** | Mann–Whitney statistics on comparisons between groups on the CFMT+. All $p$-values are unaltered, and the $\alpha$ is Bonferroni-corrected.

| Group 1 | Group 2 | $N_1$ | $N_2$ | Median 1 | Median 2 | $U$ | $p$ | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Examiners | Reviewers | 13 | 17 | 0.735 | 0.637 | 53.0 | 0.0169 | 0.0167 |
| Examiners | Students | 13 | 86 | 0.735 | 0.657 | 328.0 | 0.017 | 0.0167 |
| Reviewers | Students | 17 | 86 | 0.637 | 0.657 | 670.0 | 0.588 | 0.0167 |

## Appendix B

### B.1 | Other-Race Face Comparisons

The test created by Phillips et al. (2011) consisted of 80 pairs of face images (40 East Asian image subjects, 40 Caucasian image subjects). Participants were limited to 2 s of viewing time and responded on a 5-point scale. For our study, we expanded the viewing time to 30 s and used the same 5-point scale.

Phillips et al. (2011) used the scale to calculate the $A'$ statistic for each participant after binarizing the responses. $A'$ was used as an estimate for AUC. In our study, we used AUC for comparisons with previous studies. As AUC uses the optimal threshold, we have binarized the similarity scores ($s$) 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 B1.

We also binarized with 0, 1, and 2 being a declared match and −2 and −1 being a declared non-match. After binarizing the scores, we looked at the percent correct for each group on each set, seen in Table B2.

**TABLE B1** | Binarized group accuracy on Other-Race Face Comparisons (positive).

| Stimulus race | Examiners | Reviewers | Caucasian Students | East Asian Students |
|---|---|---|---|---|
| Caucasian | 0.905 | 0.886 | 0.824 | 0.806 |
| East Asian | 0.836 | 0.792 | 0.791 | 0.796 |

**TABLE B2** | Binarized group accuracy on Other-Race Face Comparisons (non-negative).

| Stimulus race | Examiners | Reviewers | Caucasian Students | East Asian Students |
|---|---|---|---|---|
| Caucasian | 0.914 | 0.893 | 0.824 | 0.805 |
| East Asian | 0.850 | 0.808 | 0.787 | 0.791 |

## Appendix C

### C.1 | Façade

The test created by Noyes and Jenkins (Noyes and Jenkins 2019) 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 s 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 (Noyes and Jenkins 2019) because AUC was used instead of percent correct. To compare our results to the analogous Experiment 1 in (Noyes and Jenkins 2019), we binarized the similarity scores ($s$) 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 C1.

**TABLE C1** | Binarized group accuracy on Façade (positive). ND stands for "no disguise."

| Set | Examiners | Reviewers | Caucasian Students | East Asian Students | (Noyes and Jenkins 2019) Students |
|---|---|---|---|---|---|
| ND Match | 0.960 | 0.932 | 0.905 | 0.925 | 0.950 |
| ND Non-Match | 0.964 | 0.948 | 0.889 | 0.857 | 0.920 |
| Evasion | 0.623 | 0.512 | 0.662 | 0.614 | 0.600 |
| Impersonation | 0.893 | 0.818 | 0.766 | 0.696 | 0.820 |

We also binarized with 0, 1, and 2 being a declared match and −2 and −1 being a declared non-match. After binarizing the scores, we looked at the percent correct for each group on each set, seen in Table C2.

**TABLE C2** | Binarized group accuracy on Façade (non-negative). ND stands for "no disguise."

| Set | Examiners | Reviewers | Caucasian Students | East Asian Students | (Noyes and Jenkins 2019) Students |
|---|---|---|---|---|---|
| ND Match | 0.976 | 0.938 | 0.922 | 0.933 | 0.950 |
| ND Non-Match | 0.917 | 0.926 | 0.858 | 0.806 | 0.920 |
| Evasion | 0.790 | 0.574 | 0.718 | 0.680 | 0.600 |
| Impersonation | 0.806 | 0.769 | 0.725 | 0.658 | 0.820 |

## Appendix D

### D.1 | CFMT

For better comparisons to other CFMT+ studies, we plot Figure 6 as Number Correct instead of Percent Correct (Figure D1).

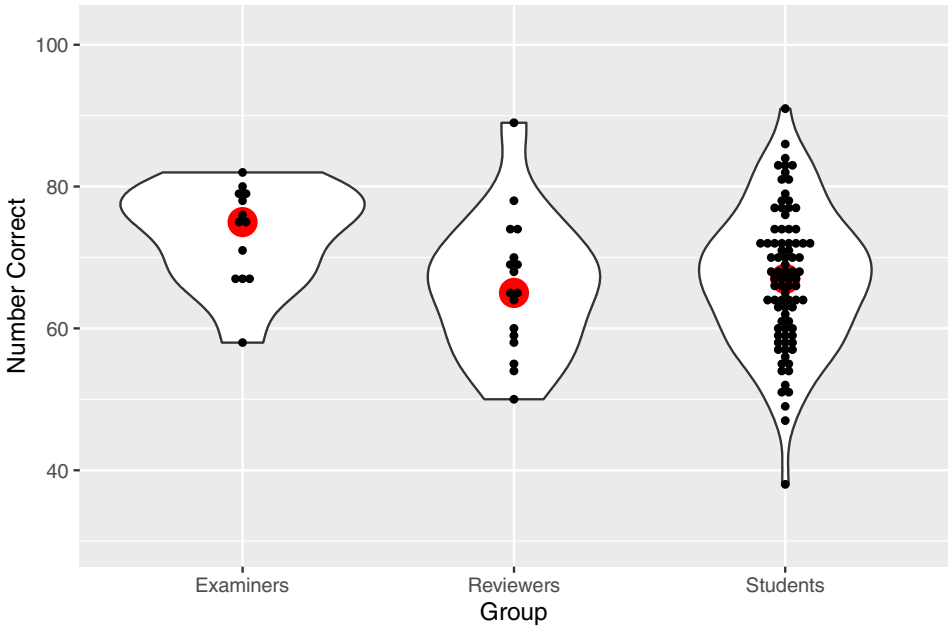



**FIGURE D1** | Group accuracy on the CFMT+. In the graph, the *x*-axis indicates the group, and the *y*-axis is number correct (maximum 102). Each black dot represents an individual participant. The violin plot shows the density. The large red dots indicate the median Number Correct for each group.

## Appendix E

### E.1 | Correlations

**TABLE E1** | The Spearman rho ($\rho$) correlations between group performance on across tests.

| | Caucasian stimuli (ORE) | East Asian stimuli (ORE) | Non-disguised condition (disguise) | Impersonation condition (disguise) | Evasion condition (disguise) |
|---|---|---|---|---|---|
| East Asian Stimuli (ORE) | Overall: 0.524<br>Examiners: 0.733<br>Reviewers: 0.430<br>C Students: 0.365<br>EA Students: 0.637 | | | | |
| Non-disguised condition (disguise) | Overall: 0.461<br>Examiners: 0.606<br>Reviewers: 0.402<br>C Students: 0.371<br>EA Students: 0.270 | Overall: 0.385<br>Examiners: 0.441<br>Reviewers: 0.212<br>C Students: 0.336<br>EA Students: 0.296 | | | |
| Impersonation condition (disguise) | Overall: 0.541<br>Examiners: 0.874<br>Reviewers: 0.313<br>C Students: 0.472<br>EA Students: 0.338 | Overall: 0.460<br>Examiners: 0.793<br>Reviewers: 0.295<br>C Students: 0.415<br>EA Students: 0.364 | Overall: 0.824<br>Examiners: 0.595<br>Reviewers: 0.631<br>C Students: 0.832<br>EA Students: 0.730 | | |
| Evasion condition (disguise) | Overall: 0.438<br>Examiners: 0.657<br>Reviewers: 0.511<br>C Students: 0.320<br>EA Students: 0.406 | Overall: 0.533<br>Examiners: 0.615<br>Reviewers: 0.647<br>C Students: 0.470<br>EA Students: 0.433 | Overall: 0.679<br>Examiners: 0.432<br>Reviewers: 0.568<br>C Students: 0.741<br>EA Students: 0.589 | Overall: 0.631<br>Examiners: 0.704<br>Reviewers: 0.176<br>C Students: 0.628<br>EA Students: 0.636 | |
| Memory (CFMT) | Overall: 0.184<br>Examiners: 0.188<br>Reviewers: 0.267<br>Students: 0.124 | Overall: 0.228<br>Examiners: 0.553<br>Reviewers: 0.037<br>Students: 0.196 | Overall: 0.331<br>Examiners: 0.124<br>Reviewers: 0.193<br>Students: 0.293 | Overall: 0.430<br>Examiners: 0.103<br>Reviewers: 0.310<br>Students: 0.409 | Overall: 0.352<br>Examiners: 0.388<br>Reviewers: 0.190<br>Students: 0.319 |

## Appendix F

### F.1 | 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 (Chang et al. 2019) 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