In press: Journal of Experimental Psychology – General
Accepted: 13th September 2022

Does a lack of perceptual expertise prevent participants from forming reliable first impressions of "other-race" faces?

Maria Tsantani[1], Harriet Over[2], Richard Cook*[1,2]

[1]Department of Psychological Sciences,
Birkbeck, University of London, London, U.K.

[2]Department of Psychology,
University of York, York, U.K.

*Correspondence:
richard.cook@bbk.ac.uk
Department of Psychological Sciences,
Birkbeck, University of London,
Malet Street,
London, U.K., WC1E 7HX

Abstract

Many studies investigating first impressions from faces employ stimulus sets that comprise only White faces. It is argued that participants lack the necessary perceptual expertise to provide reliable trait evaluations when viewing faces from ethnicities that differ from their own. In combination with a reliance on White and WEIRD participants, this concern has contributed to the widespread use of White face stimuli in this literature. The present study sought to determine whether concerns about the use of so-called “other-race” faces are justified by assessing the test-retest reliability of trait judgements made about same- and other-race faces. In two experiments conducted on 400 British participants, we find that White-British participants made reliable trait judgements about Black faces, and Black-British participants made reliable trait judgements about White faces. It is important that future work be conducted to determine how widely these results generalize. In light of our findings, however, we suggest i) that the default assumption in future first impressions research should be that participants – particularly those recruited from diverse communities – are able to form reliable first impressions of other-race faces, and ii) that faces of color are included in stimulus sets wherever possible.

Keywords:

First impressions; Trait evaluations; Face perception; Other-race effect; Cultural differences

Introduction

When we encounter someone for the first time, we spontaneously form an impression of their likely traits and characteristics (e.g., judgements about their trustworthiness, competence, and intelligence) based on their facial appearance (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Zebrowitz, 2017). First impressions are consistent across different observers even when facial stimuli are presented very briefly (Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006). While our first impressions are typically inaccurate, they can have serious real-world consequences (Olivola, Funk, & Todorov, 2014). For example, first impressions are thought to affect criminal sentencing (Wilson & Rule, 2015) and voter preferences (Todorov, Mandisodza, Goren, & Hall, 2005).

Those who participate in first impressions research typically view a series of facial images, and are asked to rate the likely traits of the people depicted. Some authors in this field use artificial computer-generated faces as stimuli (Cogsdill, Todorov, Spelke, & Banaji, 2014; Oosterhof & Todorov, 2008; Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Other authors prefer to use photographic images of real people. Some studies employ tightly controlled stimuli, where the people depicted are photographed under consistent lighting conditions, while exhibiting similar facial expressions (e.g., Cogsdill & Banaji, 2015; Talamas, Mavor, Axelsson, Sundelin, & Perrett, 2016; Willis & Todorov, 2006). Other studies use naturalistic 'ambient' images that vary widely in pose, expression and lighting conditions (e.g., Collova, Sutherland, & Rhodes, 2019; Sutherland et al., 2018; Sutherland et al., 2013).

In the majority of studies of first impressions, the stimulus sets used comprise only White faces (e.g., Cogsdill & Banaji, 2015; Collova et al., 2019; Eggleston, Flavell, Tipper, Cook, & Over, 2021; Oosterhof & Todorov, 2008; South-Palomares, Sutherland, & Young, 2018; Sutherland et al., 2013; Swe et al., 2020; Talamas et al., 2016; Todorov et al., 2013; Todorov et al., 2009; Vernon, Sutherland, Young, & Hartley, 2014; Willis & Todorov, 2006). It is not always clear why faces of color are excluded from stimulus sets (Cook & Over, 2021). Because authors so rarely explain this decision in print, readers can only speculate about their rationale. We suspect there are a number of factors to blame for this practice including the development of and adherence to research norms. Until quite recently, it was also far easier to access sets of tightly controlled White face images (with appropriate usage rights), than images of faces of color (for further discussion see Cook & Over, 2021).

Where the lack of diversity is addressed explicitly in research articles, authors cite the need to control for the so-called "other-race effect" (Collova et al., 2019; South-Palomares et al.,

2018; Sutherland et al., 2013; Swe et al., 2020; Vernon et al., 2014). The other-race effect (ORE) refers to a phenomenon whereby some individuals are better able to perceive differences between faces from their own ethnicity than faces of a different ethnicity (Furl, Phillips, & O'Toole, 2002; O'Toole & Natu, 2013; Valentine, 1991). While this rationale has not been elaborated further, a potentially valid concern is that a lack of perceptual expertise leads participants to provide unreliable (inconsistent) trait judgements. If participants were unable to provide reliable (consistent) trait judgements – i.e., if participants provided a different answer each time they judged the same face – this would make it very hard to study the resulting first impressions in a meaningful way. In principle, this situation might arise because the to-be-judged faces appear so homogenous that participants are forced to guess when asked about the traits of given target face. Alternatively, participants might form inconsistent perceptual representations of each target face and therefore provide inconsistent trait judgements.

To date, the possibility that the ORE prevents participants in first impressions research from providing reliable trait judgements of other-race faces has received little scrutiny. However, there are several reasons to question this argument. First, not everyone shows OREs. People are thought to develop expertise for the types of faces to which they are exposed (Furl et al., 2002; Sangrigoli, Pallier, Argenti, Ventureyra, & de Schonen, 2005; Valentine, 1991). For example, adults of Korean origin adopted by White families living in France showed better recognition of White faces than of East Asian faces (Sangrigoli et al., 2005). The overwhelming majority of first impressions research is conducted using participants recruited from diverse societies (U.S., U.K., Australia, France, Germany, Netherlands). It seems unjustified to routinely assume that local participants in these studies lack perceptual expertise for diverse faces. Individuals growing up in London, Paris or New York will frequently have to identify individuals (e.g., teachers, class-mates, co-workers) from a range of ethnic backgrounds. This ‘individuation experience’ is thought to be crucial for the development of perceptual expertise for faces (Richler, Wong, & Gauthier, 2011; Wong, Palmeri, & Gauthier, 2009).

Second, participants’ trait evaluations appear to depend on a relatively crude facial analysis. Individuals with developmental prosopagnosia (DP) – a neurodevelopmental disorder associated with lifelong face recognition difficulties (Cook & Biotti, 2016; Duchaine & Nakayama, 2006) – make broadly typical judgements of facial traits (Todorov & Duchaine, 2008). This condition impairs the perceptual encoding of face shape (Biotti, Gray, & Cook, 2019), disrupts the interpretation of facial emotion (Biotti & Cook, 2016), and is associated

with imprecise classification of facial sex (Marsh, Biotti, Cook, & Gray, 2019). Compared to the severe face recognition problems seen in DP, the perceptual deficits associated with the ORE are mild (Wan et al., 2017). If people with DP can make broadly typical trait evaluations, it seems unlikely that more subtle perceptual problems arising from the ORE should impair the formation of reliable first impressions.

The present study sought to test whether participants from a diverse society (the U.K.) possess sufficient perceptual expertise to form reliable impressions of likeability and intelligence when viewing other-race faces. We addressed this question in two experiments with separate samples of Black-British and White-British participants. Historically, a great deal of first impressions research has been conducted using samples of British participants (e.g., Eggleston, Flavell, et al., 2021; Eggleston, Geangu, Tipper, Cook, & Over, 2021; Ewing, Sutherland, & Willis, 2019; Kramer, Mileva, & Ritchie, 2018; Mileva, Young, Kramer, & Burton, 2019; Stirrat & Perrett, 2010; Sutherland et al., 2013; Sutherland, Oldmeadow, & Young, 2016; Talamas et al., 2016; Vernon et al., 2014). Despite the fact that Britain is an increasingly diverse society, it has been argued that British participants should not be asked to judge the traits of faces of color because of concerns about the ORE (e.g., Sutherland et al., 2013; Talamas et al., 2016; Vernon et al., 2014). We elected to use likeability and intelligence judgements because the attribution of these traits is commonly studied in first impressions research (e.g., Talamas et al., 2016; Willis & Todorov, 2006) and because they load on two dimensions thought to be crucial for social evaluation – perceived warmth and competence (Fiske, Cuddy, & Glick, 2007).

Experiment 1

In our first experiment we sought to investigate the test-retest reliability of likeability ratings made by 100 White-British and 100 Black-British participants, about 40 White and 40 Black faces. High levels of test-retest reliability when judging other-race faces would suggest that concerns about the ORE do not justify the exclusion of faces of color from first impressions research conducted on British participants. Participants also completed an Inter-Ethnicity Contact Questionnaire (IECQ) adapted from Cenac et al. (2019), which assessed participants' contact with White and Black individuals during the first 18 years of their lives. The study was conducted online using the Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020).

Methods

Transparency and openness

In the sections below we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All data and analysis code are available via the Open Science Framework (https://osf.io/hx79f/). The experimental task is available as Open Materials at gorilla.sc (https://app.gorilla.sc/openmaterials/275987). Data were analysed using Matlab, version R2021a Matlab, version R2021a (The MathWorks Inc, Natick, Massachusetts), and R, version 4.0.4 (R-Core-Team, 2021). This study's design and its analyses were not pre-registered.

Participants

Experiment 1 employed two groups of participants: 100 participants who identified as White (M_{age} = 33.95, SD_{age} = 11.27 years, 57 female, 42 male, 1 non-binary), and 100 participants who identified as Black (M_{age} = 29.50, SD_{age} = 10.45 years, 70 female, 30 male). Eleven participants in the final sample were replacements (i.e., 11 members of the original sample were excluded and replaced in order to achieve our pre-specified sample-size). Seven participants (3 White and 4 Black) were replaced due to technical problems during testing. Four participants (2 White and 2 Black) were replaced having achieved reliability scores that were lower than 2.5 SDs from their group mean. Participants were recruited through Prolific (www.prolific.co). All participants were required to be between 18 and 60 years old, to have normal or corrected-to-normal vision, to have had no clinical diagnosis of autism spectrum disorder, and to have a Prolific study approval rate of 80% or higher. Participants were required to be UK nationals currently living in the UK, and to have grown-up in the UK.

The study was approved by the Departmental Ethics Committee for Psychological Sciences, Birkbeck, University of London. The research was conducted in line with the ethical guidelines laid down in the 6th (2008) Declaration of Helsinki. Participants provided informed consent and were reimbursed for their time. Sample size was determined *a priori*. Power analysis conducted using GPower 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) revealed that a group size of 100 was sufficient to detect an effect size of 0.4 when conducting an independent samples *t*-test with a target power of 80% and alpha level of 0.05. We planned to use this analysis to compare the reliability scores for the ratings made by White and Black participants.

Display calibration

Before the experiment, participants completed a display calibration procedure, during which they were asked to adjust a rectangle until it was the same size as a credit card. This procedure has been widely used in perception research conducted online (e.g., Kramer,

Mulgrew, & Reynolds, 2018; Kramer & Reynolds, 2018). If completed correctly, this calibration procedure ensured that the stimuli in the rating task were presented at 5.5 cm × 7.0 cm irrespective of the size of monitor being used.

Face rating task

At the start of the experiment participants were provided with information about the study. We assured participants that the rationale would be fully explained at the end of the experiment. At the outset, however, participants were told that racism was not the focus of the study.

In total participants rated 40 White faces (20 men, 20 women) and 40 Black faces (20 men, 20 women). All faces were rated twice, once in the first block of 80 trials and once in the second block of 80 trials. There was a 2-minute interval in between the two blocks during which participants took a break. Presentation order was randomized across participants, but the order of trials within the first and second block was held constant for each participant. This ensured a similar interval between the first and second presentation of each face encountered during the experiment. By eliminating within-group variability attributable to order differences across the first and second block, we hoped to enhance our sensitivity to detect between-group differences.

Each trial began with a fixation cross (1 sec) followed by the face image (1 sec). Face images were obtained from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) and Shutterstock (www.shutterstock.com), and were selected at random. Face images were cropped to show only the head and neck, and aligned so that the eyes appeared in a constant position. All faces were front-facing and featured a neutral expression. Example images are shown in Figure 1.

Figure-1

The face image was replaced by a response screen, where participants rated the likeability of the person depicted using a slider that ranged from -50 (very dislikeable) to 50 (very likeable). There was no time limit for participants' responses. In the instructions provided at the start of the task, participants were told that "a likeable person would be considered trustworthy, nice, and friendly, whereas a dislikeable person would be considered untrustworthy, nasty, and unfriendly". The face rating task employed was intended to be representative of the paradigms typically used in the area.

Inter-Ethnicity Contact Questionnaire

Having completed the face rating task, participants completed the IECQ. This measure consists of six statements: 1) Most days, I encountered peers with [ethnicity] faces in educational or social contexts; 2) In my local community, many people were [ethnicity]; 3) Most days, I had face-to-face interactions with [ethnicity] people; 4) I saw many [ethnicity] individuals in TV shows, films, and online videos; 5) I saw many [ethnicity] individuals in printed media (e.g., newspapers, magazines, books); 6) Many of the characters depicted in the advertising materials I was exposed to were [ethnicity]. Participants were required to indicate the extent to which each statement described their own personal experiences with White and Black individuals, during the first 18 years of their life. Agreement was indicated on a 7-point scale ranging from disagree strongly ('1') to agree strongly ('7').

Data analysis

Test-retest reliability was calculated for each participant as the Spearman correlation between ratings of the same faces in the first and second blocks. We employed a non-parametric rank correlation to ensure that test-retest reliability scores were not inflated by extreme ratings awarded to one or two faces. Reliability scores were subjected to a Fisher z-transform prior to significance testing. This transform is necessary in order to do significance testing on correlation coefficients (the upper and lower bounds of 1.0 and 0 are removed). All reported means and standard deviations have been reverse-transformed. Cohen's *d* effect sizes and 95% confidence intervals were calculated using the rstatix package (v0.7.0; function 'cohens_d') in R. Confidence intervals were calculated by applying bootstrap resampling with 2000 replications and using the percentile interval method. All reported *p*-values are two-tailed.

Results

Test-retest reliability

Our main interest was in whether participants showed substantial test-retest reliability for both same and other-race faces. Figure 2A shows the test-retest reliabilities for likeability ratings of White and Black faces, for participants who identified as White and Black, respectively. White participants achieved mean test-retest reliabilities of .70 (*SD* = .28) for White faces, and .64 (*SD* = .28) for Black faces. Mean reliabilities for Black participants were .58 (*SD* = .28) for White faces and .62 (*SD* = .32) for Black faces. One sample *t*-tests confirmed that all four distributions significantly exceeded zero (all *t*'s > 22.0, all *p*'s < .001).

Figure-2

Reliability scores were subjected to ANOVA with Face Type (White, Black) as a within-subjects factor and Participant Ethnicity (White, Black) as a between-subjects factor. This analysis revealed a significant main effect of Participant Ethnicity [$F(1,198) = 9.718$, $p = .002$, $\eta_p^2 = .047$] and a significant Participant Ethnicity × Face Type interaction [$F(1,198) = 21.085$, $p < .001$, $\eta_p^2 = .096$]. There was no main effect of Face Type [$F(1,198) = 1.316$, $p = .253$, $\eta_p^2 = .007$]. Pairwise comparisons showed that White participants achieved higher reliabilities for White faces than Black faces [$t(99) = 4.301$, $p < .001$, $d = 0.430$, 95% CI [0.25, 0.63]], whereas Black participants achieved higher reliabilities for Black faces than White faces [$t(99) = 2.312$, $p = .023$, $d = 0.231$, 95% CI [0.04, 0.43]]. Reliabilities for White faces were significantly higher for White participants, compared with Black participants [$t(198) = 4.965$, $p < .001$, $d = 0.702$, 95% CI [0.43, 0.98]], whereas reliabilities for Black faces were similar between White and Black participants [$t(198) = .801$, $p = .424$, $d = 0.113$, 95% CI [-0.16, 0.39]]. These findings suggest that first impressions of likeability may be affected to some degree by the ORE.

Interestingly, the reliability scores achieved by White participants ($r_s = .63$, $p < .001$; Figure 3A) and Black participants ($r_s = .57$, $p < .001$; Figure 3A) when rating White and Black faces were correlated. Participants who formed reliable first impressions of White faces also tended to form reliable first impressions of Black faces, regardless of their own ethnicity.

Figure-3

In order to evaluate the pattern of ratings awarded to the 40 White and 40 Black faces by the White and Black participants, we also conducted an items analysis. Having averaged the face ratings across the two presentation blocks, we found that the mean likeability ratings awarded to the target faces by the White and Black participants were highly correlated [White faces: $r_s = .96$, $p < .001$; Black faces: $r_s = .93$, $p < .001$]. This finding indicates that the White and Black faces deemed more likeable by White participants, were also deemed more likeable by Black participants (Figure 4A).

Figure-4

Analysis of participants' IECQ scores indicated that the Black participants ($M = 6.20$, $SD = 0.79$) and White participants ($M = 6.27$, $SD = 0.84$) reported similar levels of contact with

White individuals [$t(198) = .663$, $p = .508$, $d = 0.094$, 95% CI [-0.18, 0.38]]. The Black participants ($M = 3.56$, $SD = 1.40$) reported slightly more contact with Black individuals than the White participants ($M = 3.23$ $SD = 1.27$), but this difference was not significant [$t(198) = 1.710$, $p = .089$, $d = 0.242$, 95% CI [-0.05, 0.52]]. There were no significant correlations between participants' IECQ scores and their reliability scores for White or Black faces (all $r_s < .18$). The correlations between IECQ scores and participants' reliability scores are summarized in Table 1.

Table-1

Likeability ratings

We also analyzed the mean likeability ratings awarded by White and Black participants to the White and Black faces (Figure 5A). Across the two blocks, White participants awarded a mean likeability rating of 3.47 ($SD = 9.23$) to the White faces and a mean likeability rating of 5.40 ($SD = 8.84$) to the Black faces. Black participants awarded a mean likeability rating of -0.88 ($SD = 9.62$) to the White faces and a mean likeability rating of 6.63 ($SD = 9.06$) to the Black faces. These distributions were subjected to ANOVA with Face Type (White, Black) as a within-subjects factor and Participant Ethnicity (White, Black) as a between-subjects factor.

Figure-5

The analysis revealed a main effect of Face Type [$F(1,198) = 9.718$, $p < .001$, $\eta_p^2 = .259$] and a significant Participant Ethnicity × Face Type interaction [$F(1,198) = 24.206$, $p < .001$, $\eta_p^2 = .109$]. There was no main effect of Participant Ethnicity [$F(1,198) = 1.769$, $p = .185$, $\eta_p^2 = .009$]. Black faces were rated as more likeable than White faces by White participants [$t(99) = 2.674$, $p = .009$, $d = 0.267$, 95% CI [0.06, 0.51]] and by Black participants [$t(99) = 8.579$, $p < .001$, $d = 0.858$, 95% CI [0.71, 1.04]]. White faces received higher likeability ratings by White participants than Black participants, [$t(198) = 3.262$, $p = .001$, $d = 0.461$, 95% CI [0.18, 0.74]], whereas ratings of Black faces were similar across groups [$t(198) = .977$, $p = .330$, $d = 0.138$, 95% CI [-0.15, 0.42]].

Black participants who reported greater contact with White individuals on the IECQ, tended to award higher likeability ratings to Black faces [$r_s = .23$, $p = .020$]. All other correlations with IECQ scores were non-significant (all r's < .15). The correlations between IECQ scores and mean ratings are summarized in Table 2.

Table-2

Experiment 2

In our second experiment we sought to investigate the test-retest reliability ratings of a different trait, intelligence, made by another 100 White-British and 100 Black-British participants, about the same 40 White and 40 Black faces. Once again, the study was conducted online using the Gorilla Experiment Builder (Anwyl-Irvine et al., 2020).

Methods

Transparency and openness

In the sections below we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All data and analysis code are available via the Open Science Framework (https://osf.io/hx79f/). The experimental task is available as Open Materials at gorilla.sc via (https://app.gorilla.sc/openmaterials/275987). Data were analysed using Matlab, version R2021a (The MathWorks Inc, Natick, Massachusetts), and R, version 4.0.4 (R-Core-Team, 2021). This study's design and its analyses were not pre-registered.

Participants

Experiment 2 employed two groups of participants: 100 participants who identified as White (M_{age} = 30.52, SD_{age} = 10.96 years, 60 female, 39 male, 1 non-binary), and 100 participants who identified as Black (M_{age} = 27.29, SD_{age} = 8.79 years, 75 female, 24 male, 1 non-binary). Thirteen participants in the final sample were replacements (i.e., 13 members of the original sample were excluded and replaced in order to achieve our pre-specified sample-size). Seven participants (2 White and 5 Black) were replaced due to technical problems during testing. Three participants (2 White and 1 Black) were replaced having given a rating of '0' on all trials. Three participants (1 White and 2 Black) were replaced having achieved reliability scores that were lower than 2.5 SDs from their group mean. None of the participants in Experiment 2 took part in Experiment 1.

Experimental task

With the exception of the trait being assessed, the methods were identical to those employed in Experiment 1. Participants evaluated each face for intelligence using a slider that ranged from -50 (very unintelligent) to 50 (very intelligent). In the instructions provided at the start of the task, participants were told that "an intelligent person would be considered

knowledgeable, insightful, and likely to grasp new ideas quickly, whereas an unintelligent person would be considered ignorant, foolish, and likely to grasp new ideas slowly".

Results

Test-retest reliability

Figure 2B shows the test-retest reliabilities for intelligence ratings of White and Black faces, for participants who identified as White and Black, respectively. White participants achieved mean test-retest reliabilities of .63 (*SD* = .26) for White faces, and .63 (*SD* = .27) for Black faces. Mean reliabilities for Black participants were .52 (*SD* = .33) for White faces and .56 (*SD* = .28) for Black faces. One sample *t*-tests confirmed that all four distributions significantly exceeded zero (all *t*'s > 17.0, all *p*'s < .001).

Test-retest reliability scores were subjected to ANOVA with Face Type (White, Black) as a within-subjects factor and Participant Ethnicity (White, Black) as a between-subjects factor. This analysis revealed a significant main effect of Participant Ethnicity [$F(1,198) = 13.576$, $p < .001$, $\eta_p^2 = .064$]. There was no main effect of Face Type [$F(1,198) = 1.414$, $p = .236$, $\eta_p^2 = .007$], and no Participant Ethnicity × Face Type interaction [$F(1,198) = 1.548$, $p = .215$, $\eta_p^2 = .008$]. Pairwise comparisons showed no differences in reliabilities for White and Black faces for White participants [$t(99) = 0.042$, $p = .966$, $d = 0.004$, 95% CI [-0.20, 0.20]] or for Black participants [$t(99) = 1.598$, $p = .113$, $d = 0.160$, 95% CI [-0.03, 0.38]]. White participants achieved higher reliability scores than Black participants when rating White faces [$t(186.169) = 3.658$, $p < .001$, $d = 0.517$, 95% CI [0.24, 0.80]] and Black faces [$t(198) = 2.543$, $p = .012$, $d = 0.360$, 95% CI [.09, .63]].

Once again, the reliability scores achieved by White participants ($r_s = .41$, $p < .001$; Figure 3B) and Black participants ($r_s = .44$, $p < .001$; Figure 3B) when rating White and Black faces were correlated. Participants who formed reliable first impressions of White faces also tended to form reliable first impressions of Black faces, regardless of their own ethnicity.

An items analysis revealed that the mean intelligence ratings awarded to the target faces by the White and Black participants were highly correlated [White faces: $r_s = .88$, $p < .001$; Black faces: $r_s = .92$, $p < .001$]. Once again, the pattern of ratings awarded by the White and Black participants was broadly similar: the White and Black faces deemed more intelligent by White participants, were also deemed more intelligent by Black participants (Figure 4B).

Analysis of participants' IECQ scores revealed that Black participants (M = 6.06, SD = 0.86) and White participants (M = 6.24, SD = 0.92) reported similar levels of contact with White individuals [$t(198) = 1.454$, $p = .148$, $d = 0.206$, 95% CI [-0.07, 0.49]]. However, the Black participants (M = 3.89, SD = 1.15) reported significantly more contact with Black individuals, than the White participants (M = 3.10, SD = 1.25) [$t(198) = 4.674$, $p < .001$, $d = 0.661$, 95% CI [0.38, 0.97]]. There were no significant correlations between participants' IECQ scores and their reliability scores for White or Black faces (all $r_s < .20$). The correlations between IECQ scores and participants' reliability scores are summarized in Table 1.

Intelligence ratings

Mean intelligence ratings awarded by White and Black participants to the White and Black faces are shown in Figure 5B. Across the two blocks, White participants awarded a mean intelligence rating of 3.83 (SD = 6.30) to the White faces and a mean intelligence rating of 4.10 (SD = 7.81) to the Black faces. Black participants awarded a mean intelligence rating of 3.88 (SD = 8.80) to the White faces and a mean intelligence rating of 7.89 (SD = 10.04) to the Black faces. These distributions were subjected to ANOVA with Face Type (White, Black) as a within-subjects factor and Participant Ethnicity (White, Black) as a between-subjects factor.

The analysis revealed a main effect of Face Type [$F(1,198) = 13.233$, $p < .001$, $\eta_p^2 = .063$] and a significant Participant Ethnicity × Face Type interaction [$F(1,198) = 10.045$, $p = .002$, $\eta_p^2 = .048$]. There was no main effect of Participant Ethnicity [$F(1,198) = 3.518$, $p = .062$, $\eta_p^2 = .017$]. Black participants rated the Black faces as more intelligent than the White faces [$t(99) = 4.868$, $p < .001$, $d = 0.487$, 95% CI [0.27, 0.74]], whereas White participants gave similar ratings to the Black and White faces [$t(99) = .327$, $p = .744$, $d = 0.033$, 95% CI [-0.16, 0.24]]. Black faces received higher intelligence ratings from Black participants than from White participants [$t(198) = 2.978$, $p = .003$, $d = 0.421$, 95% CI [0.15, 0.70]]. Ratings of the White faces were similar across two participant groups [$t(179.400) = .048$, $p = .962$, $d = 0.007$, 95% CI [-0.29, 0.28]].

There were no significant correlations between participants' IECQ scores and their mean intelligence ratings in Experiment 2 (all $r_s < .15$). The correlations between IECQ scores and mean ratings are summarized in Table 2.

General discussion

Many studies of first impressions from faces employ only White face stimuli. It is not always clear why faces of color are excluded from stimulus sets (Cook & Over, 2021). However, where the exclusion of faces of color is addressed explicitly, authors often cite concerns about the ORE (Collova et al., 2019; South-Palomares et al., 2018; Sutherland et al., 2013; Vernon et al., 2014). While this rationale has not been elaborated further, a potentially valid concern is that a lack of perceptual expertise leads participants to provide unreliable (inconsistent) trait judgements when viewing other-race faces. The present study sought to examine this possibility.

Rating reliability

Across two experiments we examined the reliability of trait judgements made by White-British and Black-British participants about White and Black faces. In Experiment 1, White participants made more reliable likeability judgements about White faces, and Black participants made more reliable likeability judgements about Black faces. This finding raises the possibility that some trait judgements are affected by modest OREs (Furl et al., 2002; O'Toole & Natu, 2013; Valentine, 1991). A lack of perceptual expertise may undermine participants' ability to encode the structure of other-race faces. As a result, they may be forced to guess more often about the apparent traits of the face or base their trait ratings on a more idiosyncratic percept. In Experiment 2, however, we found no evidence that the reliability of intelligence judgements was undermined by the ORE. Judgements made by White participants about White and Black faces did not differ significantly in their test-retest reliability. The same was true of the intelligence judgements made by Black participants.

These results pose an obvious question – does the modest ORE observed in Experiment 1 justify the systematic exclusion of Black faces from first impressions research conducted on British samples? We believe it does not. First, in both experiments the judgements made by British participants about other-race faces showed substantial levels of test-reliability (mean *r*'s > .50). In total, this finding was seen four times: twice in Experiment 1 and twice in Experiment 2. Given the subjective nature of trait evaluations, the number of faces judged, the way the ratings were recorded, and the fact the experiment was conducted online, the levels of reliability seen in these experiments are impressive.

Second, evidence of a significant ORE was seen only in Experiment 1 (likeability), where White-British participants' judgements of Black faces exhibited a mean test-retest reliability of .64 and Black-British participants' judgements of White faces exhibited a mean test-retest reliability of .58. The estimates of test-retest reliability seen when assessing the likeability of

other-race faces compare favorably with estimates of test-retest reliability seen when judging the intelligence of same-race faces in Experiment 2 (M_{White} = .63 to M_{Black} = .52). Presumably few authors would argue that meaningful scientific study of first impressions of intelligence is impossible because ratings exhibit this level of test-retest reliability.

Third, we found that between-group differences in rating reliability were overshadowed by substantial within-group variability. We observed a large spread of reliability scores in each sample of White and Black participants. Importantly, in both experiments, people who formed reliable first impressions about White faces, tended to form reliable first impressions about Black faces, irrespective of their own ethnicity. This finding suggests that, in many cases, identifying unreliable raters may be a more pressing problem for first impressions research than any potential noise introduced by the ORE.

In both experiments, participants completed a self-report measure of their contact with White and Black individuals during the first 18 years of their lives. The responses revealed that both groups had similar and extensive contact with White individuals, that greatly exceeded their contact with Black individuals. In both experiments, Black-British participants reported more contact with Black individuals than White-British participants, however this difference only reached significance in Experiment 2. We observed no relationship between contact with White and Black individuals during development and participants' ability to make reliable trait judgements about White and Black faces.

The lack of a relationship between inter-group contact and judgment reliability is noteworthy because it suggests that participants from less diverse regions of the U.K. (e.g., those from rural communities), or those who grew-up in the U.K. when it was a less-diverse society, were able to make reliable trait judgements about other-race faces. This finding potentially accords with the view that trait judgements often depend on a relatively crude perceptual analysis (Cook & Over, 2021). Typically, our White participants had less contact with Black individuals than White individuals during development. Nevertheless, the clear majority had sufficient perceptual expertise to form reliable first impressions of likeability and intelligence when viewing Black faces.

In both experiments, Black participants produced slightly less reliable trait ratings than the White participants. This effect is unlikely to be a product of perceptual expertise, as it was seen irrespective of target face ethnicity. One possible explanation is that, because Black-British participants face substantial prejudice themselves (e.g., Olusoga, 2016), they are

more willing to question their appearance-based assumptions about the likely traits of strangers. It might be interesting to examine whether other victims of systematic prejudice (e.g., victims of anti-Semitism or Islamophobia) are also less consistent in their first impressions. A second possibility is that there were systematic differences in the testing environments. On average, the White participants may have used monitors of higher-quality and may have had access to quieter testing environments, than the Black participants. Again, it might be interesting to see whether a similar group difference is seen in lab-based research.

The present study sought to determine whether British participants form reliable first impressions when tested using a paradigm representative of the those typically employed in this field. However, some readers might query whether our estimates of test-retest reliability have been artificially inflated by the use of a trait-rating task that was unduly easy. We do not believe that the demands of our task were unusually light. For example, we employed a set of 80 tightly-controlled face stimuli chosen at random from existing databases. Participants viewed each face for only 1 sec and made their trait judgements after stimulus offset. The demands of the task could have been reduced considerably had we let participants view each face for longer, let them enter their ratings while the face was visible, or used stimuli that accentuated trait-relevant cues (e.g., differences in expression).

Group differences

A great deal has been written about the consistency of first impressions across observers (Todorov et al., 2015; Zebrowitz, 2017). In particular, the suggestion that some first impressions are ‘culturally universal’ has been cited as evidence that these trait judgements have an innate basis (Sutherland et al., 2020; Sutherland et al., 2018; Zebrowitz et al., 2012; Zebrowitz & Zhang, 2011). Critically, however, the focus on White face stimuli and White and WEIRD participants has likely exaggerated the extent of the inter-rater consensus (Cook & Over, 2021; Over, Eggleston, & Cook, 2020). The literature on inter-group bias includes overwhelming evidence that perceived ethnic groupings are associated with culturally acquired stereotypes, and that endorsement of these stereotypes varies as a function of group membership (Brown Givens & Monahan, 2005; Fiske et al., 2007; Fiske, Cuddy, Glick, & Xu, 2002). Consequently, the use of diverse raters and diverse face stimuli will likely reveal more heterogeneous first impressions (Cook & Over, 2021).

The effects of cultural stereotyping on first impressions were evident in our data. For example, in Experiment 1, White participants judged the set of White faces to be more

likable than did Black participants. Similarly, in Experiment 2 Black participants judged Black faces to be more intelligent than did White participants. Similar effects have been described elsewhere (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011; Xie, Flake, & Hehman, 2019; Xie, Flake, Stolier, Freeman, & Hehman, 2021; Zebrowitz, Montepare, & Lee, 1993). For example, White American participants judged Black faces to be more dominant than White or Korean faces, while Black American and Korean participants judged Black faces to be less dominant than White or Korean faces (Zebrowitz et al., 1993). Similarly, individuals who exhibit a pro-White implicit bias are likely to judge Black faces as less trustworthy than White faces, while those with a pro-Black bias show the opposite pattern (Stanley et al., 2011).

Whereas evidence of cultural universality might suggest that first impressions have an innate origin, evidence of individual differences and cross-cultural variability accords with the view that first impressions are heavily influenced by cultural learning. According to the Trait Inference Mapping (TIM) account (Cook, Eggleston, & Over, 2022; Cook & Over, 2020; Over & Cook, 2018; Over et al., 2020), first impressions are the products of associative mappings between points in face-space (representations of facial structure) and points in trait-space (representations of the potential trait profiles that others may possess). Associative mappings are thought to be acquired through correlated face-trait experience; for example, exposure to cultural instruments (e.g., propaganda, illustrated story books, art and iconography) that repeatedly pair certain types of face (e.g., handsome, square jaw, perfect smile) with particular traits (e.g., bravery, honesty, leadership). Where different individuals are exposed to different sources of correlated face-trait experience (e.g., different propaganda, different story books, different art and iconography) they would be expected to acquire different face-trait mappings (Cook et al., 2022; Cook & Over, 2020; Over & Cook, 2018; Over et al., 2020).

Directions for future research

The results of the present study suggest several avenues for future research. First, we must seek to understand how well our findings generalize to other observer groups and populations. In the present study, we found evidence that White and Black British participants form reliable first impressions of other-race (White and Black) faces. It is important to confirm that this is also true of British participants of other ethnicities (e.g., Asian-British, Arab-British). It will also be important to assess whether our findings – obtained with British participants – replicate in samples recruited from other countries. In particular, a great deal of first impressions research is conducted in the U.S., Australia, and central Europe (France, Belgium, Germany, Netherlands). We predict that samples drawn

from these diverse communities and cultures will also be able to form reliable first impressions of other-race faces when tested under conditions similar to those employed here.

Second, it will be important to ascertain whether these findings generalize to other paradigms used in first impressions research. As described above, the present study sought to determine whether British participants form reliable first impressions when tested using a paradigm representative of the those typically employed in this field. As such the perceptual demands of our task are moderate. It is possible, however, that small differences in perceptual expertise may impact performance to a greater degree on tasks with high perceptual demands. For example, in our task we presented unmanipulated images of people for 1 sec. It may be harder to form reliable impressions of other-race faces when image morphing is used to produce homogenous stimulus sets that contain only subtle variation (e.g., FeldmanHall et al., 2018). Similarly, participants may form less reliable impressions of other-race faces when presented very briefly – say, for 100 ms or less (e.g., Todorov et al., 2009; Willis & Todorov, 2006).

Third, it is important to determine whether our findings generalize to all trait judgements. Individuals who struggle to form accurate perceptual descriptions of faces may experience particular difficulties basing perceptual decisions on information from the eye-region (DeGutis, Cohan, Mercado, Wilmer, & Nakayama, 2012; Fisher, Towler, & Eimer, 2016; Tsantani, Gray, & Cook, 2022). First impressions of intelligence and likeability are thought to be based on perceptual evidence accumulated from the whole face; i.e., from the mouth, the nose, the eyes, the configuration of these features, and head shape (Oosterhof & Todorov, 2008; Talamas et al., 2016; Vernon et al., 2014). It is possible, however, that the inference of certain traits may be based disproportionately on perceptual evidence gathered from the eye region. Such impressions might be particularly susceptible to OREs.

Implications for first impressions research

Historically, the vast majority of first impressions research has been conducted using White and WEIRD participants (Cook & Over, 2021; Over et al., 2020). Many of these studies have used only White face stimuli, based – seemingly – on the untested assumption that participants may struggle to provide reliable ratings of other-race faces due to a lack of perceptual expertise (Collova et al., 2019; South-Palomares et al., 2018; Sutherland et al., 2013; Swe et al., 2020; Vernon et al., 2014). The current results reveal that this assumption

is unsafe – under the kinds of viewing conditions that are common in this literature, many participants do form reliable first impressions of other-race faces.

As we acknowledge above, there are many outstanding questions about the generalizability of these findings. Participants' ability to form reliable impressions of other-race faces may vary as a function of the viewing conditions, the particular participant groups being tested, the kinds of stimuli being judged, and the traits being inferred. As such, future research may identify certain situations in which our conclusions do not apply. In light of our findings, however, we suggest that future first impressions research should adopt a new default assumption; that participants can form reliable first impressions of other-race faces, particularly when samples are recruited from diverse communities (e.g., U.K., U.S., Australia, France, Belgium, Germany, Netherlands).

Where authors still have concerns about the influence of the ORE (e.g., because of the particular viewing conditions or stimuli employed), it is easy enough to elicit two sets of trait evaluations from each participant and thereby examine judgement reliability on a participant-by-participant basis. Indeed, a greater focus on rating reliability might benefit this field in general. Our findings suggest that rating reliability varies enormously within groups of participants who share the same ethnicity. By assuming i) that judgements of other-race faces will be unreliable, and ii) that judgements of same-race faces will be reliable, researchers may be committing two errors.

The wider case for greater diversity

The failure to include faces of color in stimulus sets has had detrimental consequences for the scientific investigation of first impressions. The lack of diversity has inflated estimates of inter-rater agreement and obscured evidence of systematic individual and cultural differences (Cook & Over, 2021; Over et al., 2020). In turn, misleading claims about cultural universality have hampered efforts to understand the origins of first impressions (Cook & Over, 2021; Over et al., 2020). The lack of diversity may have also hindered efforts to understand the accuracy of first impressions – to what extent the traits we attribute to others correspond to their actual characteristics and behaviors (Cook & Over, 2021). Similarly, this convention may have yielded misleading conclusions about the first impressions people form outside the lab and the kinds of cues they base their judgements on (Cook & Over, 2021).

Increasing the diversity present within stimulus sets may change the way people make trait evaluations (e.g., some features may be rendered more salient, while other features may

appear less salient) and some of the results may be complex (e.g., there may be evidence of contextual variability, systematic individual differences, and some feature-trait relationships may not generalize to all face types). However, this is not a problem with the approach *per se*; rather, this simply reflects the reality of the phenomenon (Cook & Over, 2021). We do not live in an all-White world where we encounter long uninterrupted runs of White faces. Our societies are increasingly diverse and the first impressions we form outside the lab arise in this context.

The prevailing focus on impressions of White faces not only undermines scientific efforts to understand the phenomenon, but it may have wider societal consequences (Cook & Over, 2021). If first impressions research fails to use more diverse face stimuli, there is a danger that researchers will inadvertently reinforce the idea that White faces are somehow 'the standard' or 'more important'. This is a particular risk when authors use White face stimuli exclusively, but generalize their conclusions to all faces / all observers, without appropriate qualifications. When reading the existing literature, it is easy to forget that people who identify as White represent a minority of the global population.

In WEIRD societies around the world, people of color face systematic discrimination in employment and criminal justice settings and in the political sphere (Pager & Shepherd, 2008; Paluck, Porat, Clark, & Green, 2021). The first impressions formed about people of color often have life-changing – and too often, fatal – consequences (Correll, Park, Judd, & Wittenbrink, 2002; Viglione, Hannon, & DeFina, 2011). In the future, there may be opportunities to use knowledge gained from the study of first impressions to combat racial bias. It would be a terrible shame if these opportunities were missed because of ill-conceived choices about stimuli (Cook & Over, 2021).

Conclusion

In summary, concerns about the ORE (Collova et al., 2019; South-Palomares et al., 2018; Sutherland et al., 2013; Vernon et al., 2014), in combination with a reliance on White and WEIRD participants, have contributed to the widespread use of White face stimuli in the first impressions literature. Our results suggest that concerns about the use of other-race faces may have been over-stated. In two experiments, we find that White-British participants make reliable trait judgements about Black faces, and Black-British participants make reliable trait judgements about White faces. Between-group differences in rating reliability were overshadowed by substantial within-group variability. It is important that future work be conducted to determine how widely these results generalize. In light of our findings,

however, we suggest i) that the default assumption in future first impressions research should be that participants – particularly those recruited from diverse communities – are able to form reliable first impressions of other-race faces, and ii) that faces of color are included in stimulus sets wherever possible.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388-407.

Biotti, F., & Cook, R. (2016). Impaired perception of facial emotion in developmental prosopagnosia. *Cortex, 81*, 126-136.

Biotti, F., Gray, K. L. H., & Cook, R. (2019). Is developmental prosopagnosia best characterised as an apperceptive or mnemonic condition? *Neuropsychologia, 124*, 285-298.

Brown Givens, S. M., & Monahan, J. L. (2005). Priming mammies, jezebels, and other controlling images: An examination of the influence of mediated stereotypes on perceptions of an African American woman. *Media Psychology, 7*(1), 87-106.

Cenac, Z., Biotti, F., Gray, K. L. H., & Cook, R. (2019). Does developmental prosopagnosia impair identification of other-ethnicity faces? *Cortex, 119*, 12-19.

Cogsdill, E. J., & Banaji, M. R. (2015). Face-trait inferences show robust child–adult agreement: Evidence from three types of faces. *Journal of Experimental Social Psychology, 60*, 150-156.

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science, 25*(5), 1132-1139.

Collova, J. R., Sutherland, C. A., & Rhodes, G. (2019). Testing the functional basis of first impressions: Dimensions for children's faces are not the same as for adults' faces. *Journal of Personality and Social Psychology*.

Cook, R., & Biotti, F. (2016). Developmental prosopagnosia. *Current Biology, 26*(8), R312-R313.

Cook, R., Eggleston, A., & Over, H. (2022). The cultural learning account of first impressions. *Trends in Cognitive Sciences, 26*(8), 656-668.

Cook, R., & Over, H. (2020). A learning model can explain both shared and idiosyncratic first impressions from faces. *Proceedings of the National Academy of Sciences of the USA, 117*(28), 16112-16113.

Cook, R., & Over, H. (2021). Why is the literature on first impressions so focused on White faces? *Royal Society Open Science, 8*(9), e211146.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314-1329.

DeGutis, J., Cohan, S., Mercado, R. J., Wilmer, J., & Nakayama, K. (2012). Holistic processing of the mouth but not the eyes in developmental prosopagnosia. *Cognitive Neuropsychology, 29*(5-6), 419-446.

Duchaine, B., & Nakayama, K. (2006). Developmental prosopagnosia: a window to content-specific face processing. *Current Opinion in Neurobiology, 16*, 166-173.

Eggleston, A., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Culturally learned first impressions occur rapidly and automatically and emerge early in development. *Developmental Science, 24*(2), e13021.

Eggleston, A., Geangu, E., Tipper, S. P., Cook, R., & Over, H. (2021). Young children learn first impressions of faces through social referencing. *Scientific Reports, 11*(1), 14744.

Ewing, L., Sutherland, C. A., & Willis, M. L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology*.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160.

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America, 115*(7), E1690-E1697.

Fisher, K., Towler, J., & Eimer, M. (2016). Reduced sensitivity to contrast signals from the eye region in developmental prosopagnosia. *Cortex, 81*, 64-78.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878-902.

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*(6), 797-815.

Kramer, R. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PloS One, 13*(8), e0202655.

Kramer, R. S., Mulgrew, J., & Reynolds, M. G. (2018). Unfamiliar face matching with photographs of infants and children. *PeerJ, 6*, e5010.

Kramer, R. S., & Reynolds, M. G. (2018). Unfamiliar face matching with frontal and profile views. *Perception, 47*(4), 414-431.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122-1135.

Marsh, J. E., Biotti, F., Cook, R., & Gray, K. L. H. (2019). The discrimination of facial sex in developmental prosopagnosia. *Scientific Reports, 9*, 19079.

Mileva, M., Young, A. W., Kramer, R. S., & Burton, A. M. (2019). Understanding facial impressions between and within identities. *Cognition, 190*, 184-198.

O'Toole, A. J., & Natu, V. (2013). Computational perspectives on the other-race effect. *Visual Cognition, 21*(9-10), 1121-1137.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18*, 566-570.

Olusoga, D. (2016). *Black and British: A forgotten history*: Pan Macmillan.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA, 105*, 11087-11092.

Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition, 170*, 190-200.

Over, H., Eggleston, A., & Cook, R. (2020). Ritual and the origins of first impressions. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 375*(1805), e20190435.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*, 181-209.

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology, 72*.

R-Core-Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* http://www.R-project.org/.

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science, 20*(2), 129-134.

Sangrigoli, S., Pallier, C., Argenti, A. M., Ventureyra, V. A., & de Schonen, S. (2005). Reversibility of the other-race effect in face recognition during childhood. *Psychological Science, 16*(6), 440-444.

South-Palomares, J. K., Sutherland, C. A., & Young, A. W. (2018). Facial first impressions and partner preference models: Comparable or distinct underlying structures? *British Journal of Psychology, 109*(3), 538-563.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108*(19), 7710-7715.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science, 21*(3), 349-354.

Sutherland, C. A., Collova, J. R., Palermo, R., Germine, L., Rhodes, G., Blokland, G. A., . . . Wilmer, J. B. (2020). Reply to Cook and Over: Social learning and evolutionary mechanisms are not mutually exclusive. *Proceedings of the National Academy of Sciences of the USA, 117*(28), 16114-16115.

Sutherland, C. A., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin, 44*, 521– 537.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*, 105-118.

Sutherland, C. A., Oldmeadow, J. A., & Young, A. W. (2016). Integrating social and facial models of person perception: Converging and diverging dimensions. *Cognition, 157*, 257-267.

Swe, D. C., Palermo, R., Gwinn, O. S., Rhodes, G., Neumann, M., Payart, S., & Sutherland, C. A. (2020). An objective and reliable electrophysiological marker for implicit trustworthiness perception. *Social Cognitive and Affective Neuroscience, 15*(3), 337-346.

Talamas, S. N., Mavor, K. I., Axelsson, J., Sundelin, T., & Perrett, D. I. (2016). Eyelid-openness and mouth curvature influence perceived intelligence beyond attractiveness. *Journal of Experimental Psychology: General, 145*(5), 603-620.

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion, 13*, 724-738.

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology, 25*(3), 395-410.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626.

Todorov, A., Olivola, C., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519-545.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*, 813-833.

Tsantani, M., Gray, K. L. H., & Cook, R. (2022). New evidence of impaired expression recognition in developmental prosopagnosia. *Cortex*.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43*(2), 161-204.

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences of the USA, 111*(32), E3353-E3361.

Viglione, J., Hannon, L., & DeFina, R. (2011). The impact of light skin on prison time for black female offenders. *The Social Science Journal, 48*(1), 250-258.

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

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325–1331.

Wong, A. C. N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: Becoming a Ziggerin expert—but which type? *Psychological Science, 20*(9), 1108-1117.

Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology, 117*(2), 364-385.

Xie, S. Y., Flake, J. K., Stolier, R. M., Freeman, J. B., & Hehman, E. (2021). Facial impressions are predicted by the structure of group stereotypes. *Psychological Science, 32*(12), 1979-1993.

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science, 26*, 237-242.

Zebrowitz, L. A., Montepare, J. M., & Lee, H. K. (1993). They don't all look alike: Individual impressions of other racial groups. *Journal of Personality and Social Psychology, 65*(1), 85-101.

Zebrowitz, L. A., Wang, R., Bronstad, P. M., Eisenberg, D., Undurraga, E., Reyes-García, V., & Godoy, R. (2012). First impressions from faces among US and culturally isolated Tsimane' people in the Bolivian rainforest. *Journal of Cross-Cultural Psychology, 43*, 119-134.

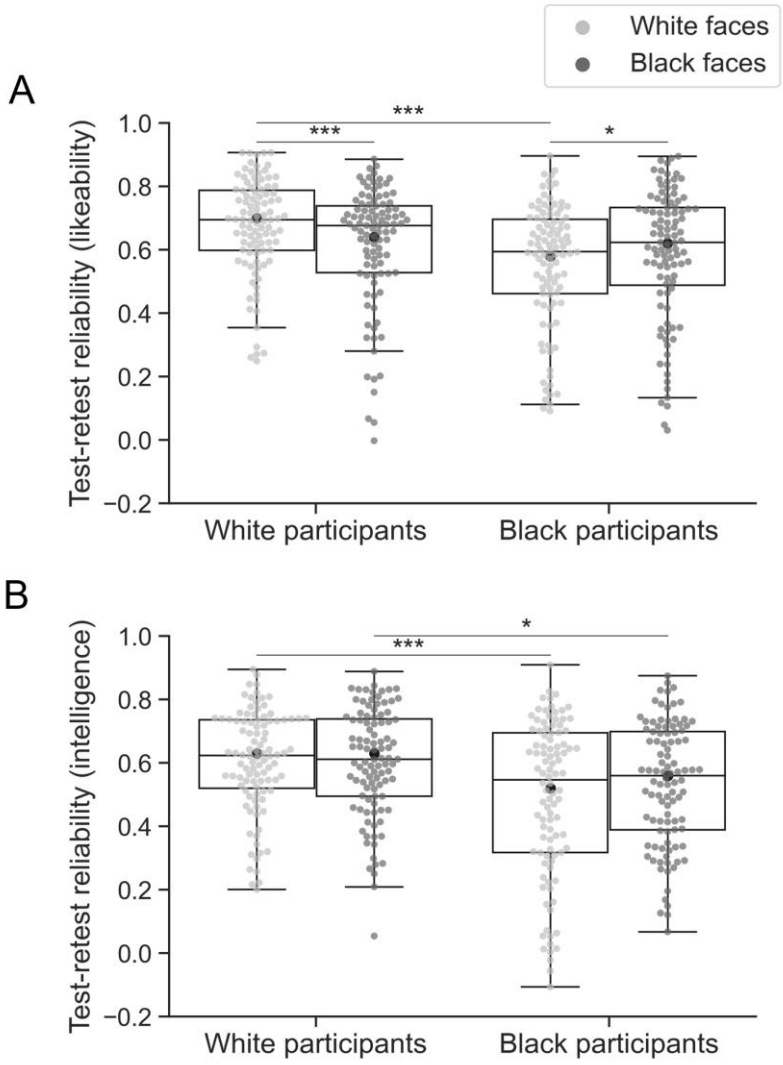
Zebrowitz, L. A., & Zhang, Y. (2011). Origins of impression formation in animal and infant face perception. In D. J. Cacioppo (Ed.), *The Handbook of Social Neuroscience* (pp. 434-444). Oxford: Oxford University Press.

Figures

Figure 1. *Examples of face stimuli used in the trait rating task.*

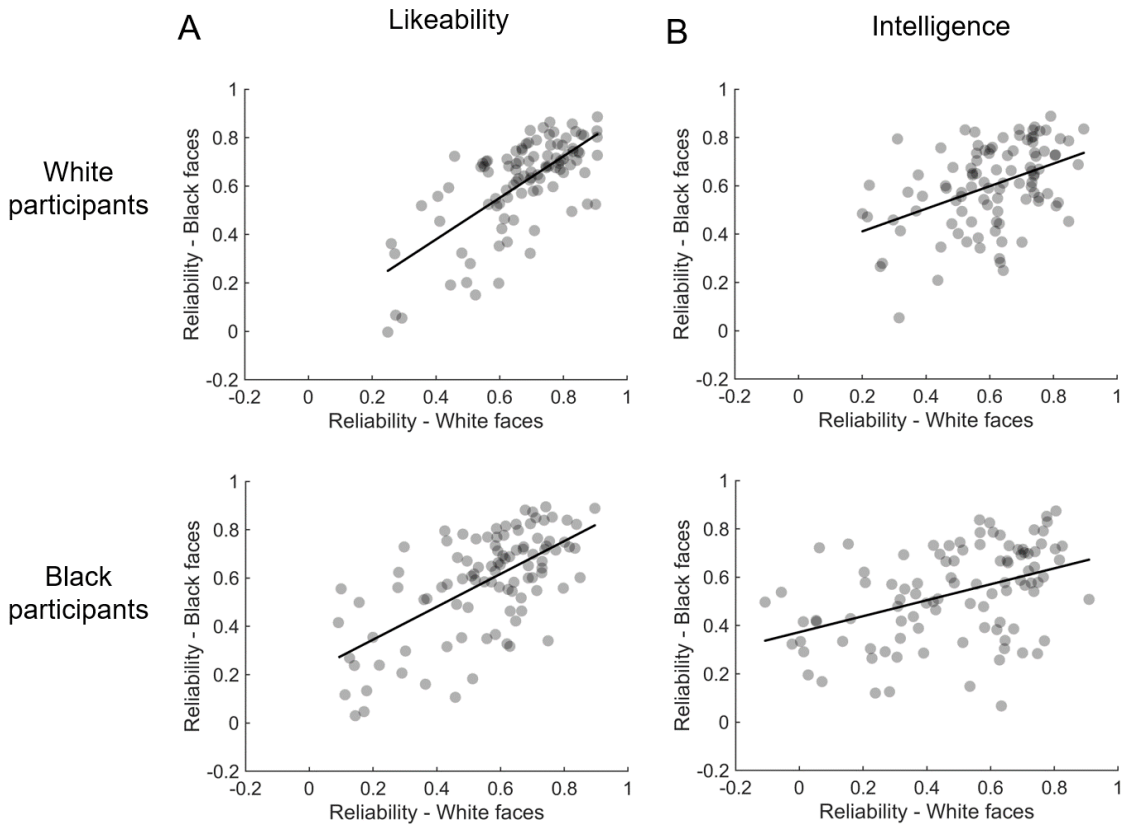
Note. Stimulus images were obtained from the Chicago Face Database (Ma et al., 2015). Images are freely available for non-commercial research purposes (https://chicagofaces.org/).

Figure 2. *Distributions of test-retest reliability scores*



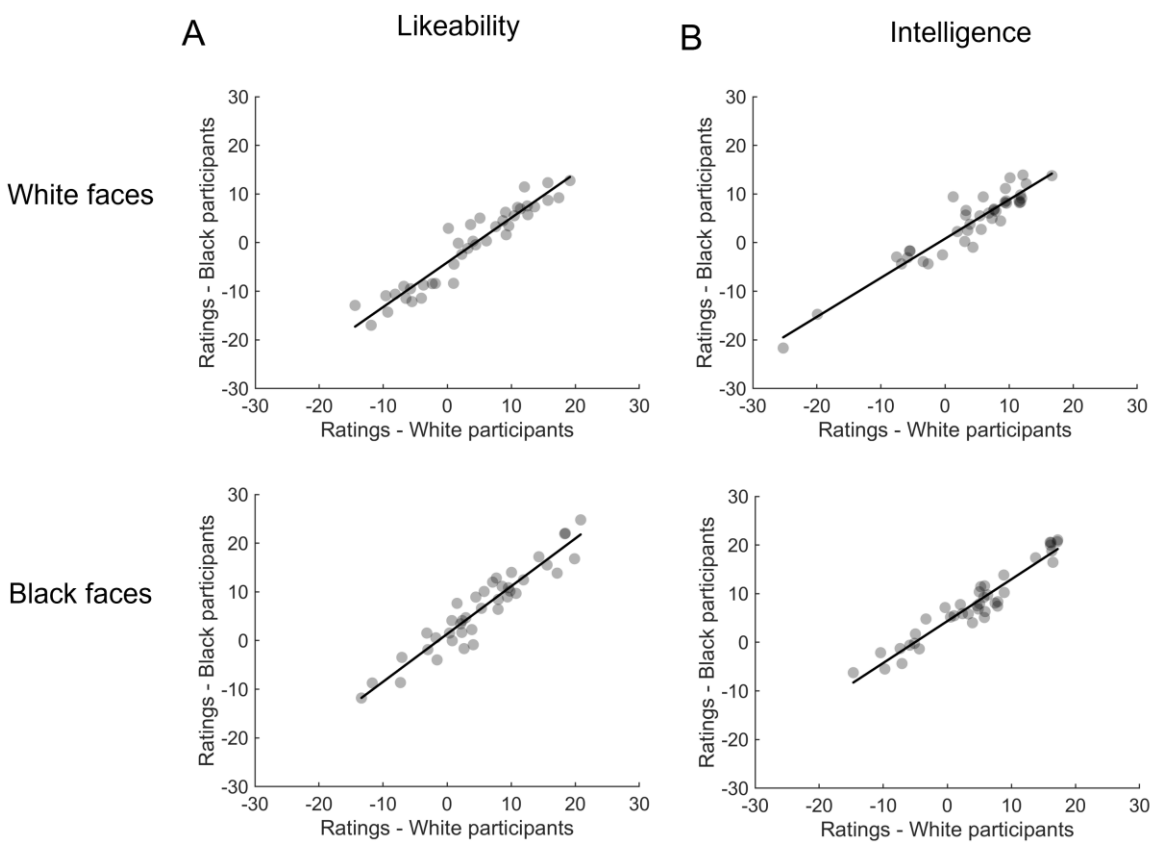
Note. Test-retest reliability (r_s) for ratings of White and Black faces on likeability (Experiment 1) (A) and intelligence (Experiment 2) (B), for White and Black participants. Boxes show the median and interquartile range, and filled black circles show the mean. The whiskers indicate the rest of the distribution, except for points that are determined to be outliers. Asterisks and vertical lines above the boxes indicate significant pairwise contrasts. * $p < .05$, *** $p < .001$

Figure 3. *Correlation of reliability scores seen for judgements of White and Black faces.*



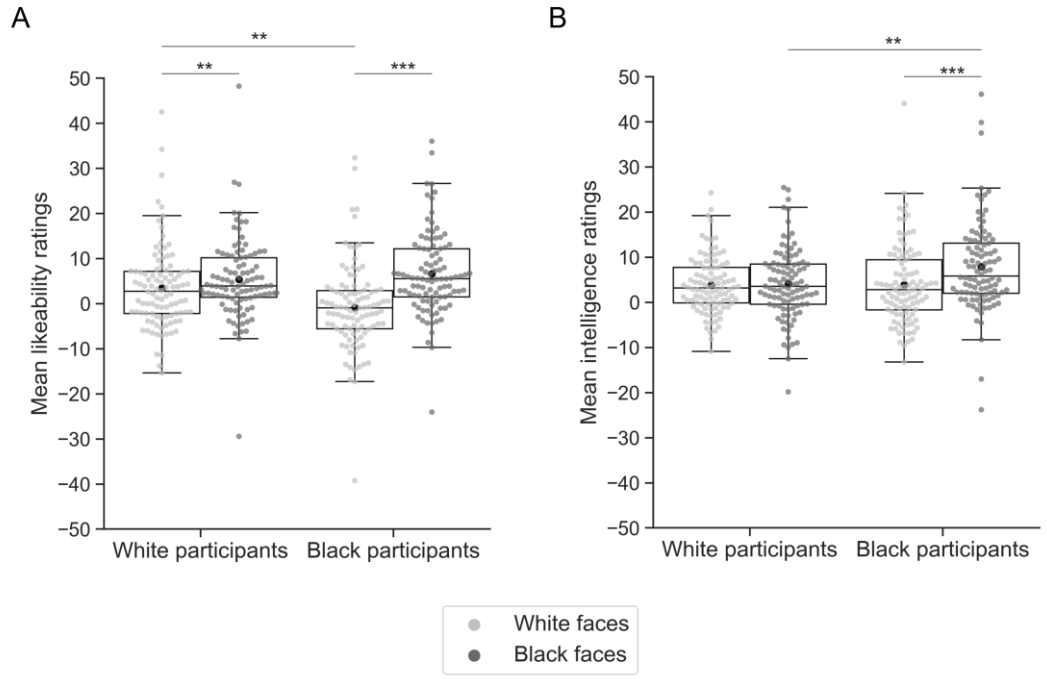
Note. Test-retest reliability scores (r_s) for White faces plotted against scores for Black faces for ratings of likeability (Experiment 1) (A) and intelligence (Experiment 2) (B), for White and Black participants. Lines show simple linear regression models.

Figure 4. *Results of the items-analyses.*



Note. (A) In Experiment 1, the White and Black faces deemed more likeable by White participants, were also deemed more likeable by Black participants. (B) In Experiment 2, the White and Black faces deemed more intelligent by White participants were also deemed more intelligent by Black participants. Lines show simple linear regression models.

Figure 5. *Mean likability and intelligence ratings*



Note. Mean ratings awarded to Black and White faces in (A) Experiment 1 and (B) Experiment 2. Boxes show the median and interquartile range, and filled black circles show the mean of each distribution. The whiskers indicate the rest of the distribution, except for points that are determined to be outliers. Asterisks and vertical lines above the boxes indicate significant pairwise contrasts. * $p < .05$, ** $p < .01$, *** $p < .001$.

Tables

Table 1. Summary of the correlations seen between IECQ scores and the reliability scores seen in Experiment 1 (likeability) and Experiment 2 (intelligence).

		White participants		**Black participants**	
		White contact	*Black contact*	*White contact*	*Black contact*
Experiment 1	White faces	.18	-.19	.08	.02
	Black faces	.09	-.09	.17	-.05
Experiment 2	White faces	.06	.02	.11	-.09
	Black faces	.19	-.07	.09	-.04

Table 2. Summary of the correlations seen between IECQ scores and the mean ratings seen in Experiment 1 (likeability) and Experiment 2 (intelligence). * p = .020

		White participants		**Black participants**	
		White contact	*Black contact*	*White contact*	*Black contact*
Experiment 1	White faces	-.07	-.03	-.12	-.00
	Black faces	.07	.12	.23*	-.09
Experiment 2	White faces	.02	.08	-.03	.07
	Black faces	-.00	-.12	.02	.07