Social Cognition, Vol. 37, No. 4, 2019, pp. 357–388

FACE-TRAIT AND FACE-RACE CUES IN ADULTS' AND CHILDREN'S SOCIAL EVALUATIONS

Tessa E. S. Charlesworth and Mahzarin R. Banaji
Harvard University

When making character judgments from faces, perceivers must integrate and prioritize a myriad of information, including perceived traits (e.g., appearance of trustworthiness, submissiveness, competence) and social category membership (e.g., Afrocentric, Eurocentric appearance). Across four studies, adults (Studies 1–3) and children (5–13 years old; Study 4) made evaluations based predominantly on face-traits. Regardless of whether face pairs were White-White, Black-Black, or White-Black, participants selected the trustworthy, submissive, or competent-appearing face as "nice." Further, although face-traits were used at all ages, face-race cues were used only by older children and adults, in line with correction for race bias. Indeed, adults' use of face-race cues decreased when motivation or ability to control race bias was reduced through time constraints or indirect responding. These findings reveal the processes underlying the prioritization of face-trait over face-race cues and provide the first developmental examination of how perceivers integrate such cues in explicit character evaluations.

Keywords: face inference, trait inference, social desirability, social cognitive development

When making character judgments from faces, social perceivers encounter a complex task: They must integrate and prioritize a myriad of information about social targets, including a target's perceived character traits and racial group member-

Supplemental materials are available online.

Parts of the data reported in this article were presented at the Biennial Conference of the Cognitive Development Society, Portland, Oregon, in October 2017. The R code and data included in this article are available at the Open Science Framework (https://osf.io/xdt7y/). We thank Benedek Kurdi and Larisa Heiphetz for comments on earlier drafts of this manuscript, and the Harvard Dean's Competitive Fund for Promising Scholarship awarded to Mahzarin R. Banaji.

The experiments in this article earned Open Materials and Open Data badges for transparent practices. Materials and data for the experiments are available at https://osf.io/xdt7y/

We thank the schools, parents, and families who participated in this research, and Wesley Cash, Maribelle Dickins, Paige Guarino, Mahima Sindhu, and Ashely Xu who assisted with data collection.

Address correspondence to Tessa Charlesworth, Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138; E-mail: tet371@g.harvard.edu

ship. The influence and relative importance of face-trait and face-race cues presents an informative comparison because each cue has features that paradoxically predict both over- and under-prioritization in social perception. On the one hand, face-trait and face-race information are used as early as infancy (Jessen & Grossmann, 2016; Kelly et al., 2005) and have pervasive consequences in real-world domains ranging from election outcomes (Payne et al., 2010; Todorov, Mandisodza, Goren, & Hall, 2005) to judgments of criminality (Pager, 2003; Wilson & Rule, 2015). On the other hand, face-trait and face-race cues are unreliable predictors of actual behavior and, in many cases, are used for inaccurate profiling or biased judgments of character (e.g., Glaser, 2015; Olivola, Funk, & Todorov, 2014). Additionally, many perceivers are motivated to appear unprejudiced about race (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002) and may therefore suppress the influence of face-race cues in social judgments.

These competing features of being, at once, early-emerging and pervasive yet unreliable and prone to correction suggest that neither face-trait nor face-race will be prioritized across all situations. As such, it is possible to identify experimental manipulations and developmental changes that sway participants towards using one cue or the other, ultimately yielding insights into when and how the fundamental features of facial appearance and race guide social judgments.

FACE-TRAIT CUES INFLUENCE SOCIAL EVALUATIONS

Sensitivity to the appearance of face-traits does not require prolonged experience in the social world: As early as 7 months of age, infants prefer to look at trustworthy faces over neutral or untrustworthy faces (Jessen & Grossmann, 2016). Children at 3 years of age make explicit global judgments of "niceness" on the basis of facial trustworthiness, submissiveness, and competence (Cogsdill, Todorov, Spelke, & Banaji, 2014), and by 5 years of age children's face-trait judgments are indistinguishable from the judgments of adults, even for naturalistic faces and faces of other species (Cogsdill & Banaji, 2015). Moreover, face-trait cues exert their influence relatively automatically (Willis & Todorov, 2006), are largely consistent across cultures (Rule et al., 2010), and predict consequential real-world outcomes (for a review, see Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). For these reasons, face-trait cues would be expected to guide explicit social evaluations.

However, there is no evidence that face-trait cues are accurate predictors of a target's character (Olivola et al., 2014; Todorov et al., 2015), and perceivers may therefore negate face-trait information when it is deemed less relevant or predictive for their social judgments. For instance, face-trait cues of competence may be perceived to be particularly irrelevant to judgments of warmth or niceness because competence and warmth are argued to be orthogonal dimensions of social evaluations (Fiske, Cuddy, Glick, & Xu, 2002). Including face-trait variations of competence, alongside trustworthiness and dominance, therefore provides an informative test of possible limits on the use of face-trait cues in social evaluations.

Additionally, compared to the ease of categorizing by race, discriminating face-trait cues is more difficult: Even adults categorizing two White faces on face-traits of competence, submissiveness, or trustworthiness do not reach perfect categorizations (Todorov, Said, Engell, & Oosterhof, 2008). Further, a strong interpretation of the "other race effect" (Meissner & Brigham, 2001) would suggest that difficulty with face-trait cue discriminability will increase when the faces are from a less-familiar outgroup. This difficult discriminability, as well as the perceived irrelevance or inaccuracy, would suggest that face-trait cues may not be used in social evaluations, especially when other cues (e.g., face-race) are available.

FACE-RACE CUES INFLUENCE SOCIAL EVALUATIONS

Like face-trait cues, sensitivity to face-race cues emerges as early as infancy (Bar-Haim, Ziv, Lamy, & Hodes, 2006; Kelly et al., 2005). By at least 4 years of age, children explicitly differentiate and evaluate racial groups (Aboud, 1988; Clark & Clark, 1947; Raabe & Beelmann, 2011), and by 6 years of age, children show adult-like implicit associations of White faces with positive stimuli and Black faces with negative stimuli (Baron & Banaji, 2006). Racial group membership also has pervasive consequences for prejudice and discrimination: Implicit and explicit racial prejudice is apparent across millions of respondents in the U.S. and abroad (Charlesworth & Banaji, 2019; Nosek et al., 2007) and has ramifications for racial disparities in domains, including policing (Hehman, Flake, & Calanchini, 2017), medicine (Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016), hiring (Pager, 2003), and politics (Payne et al., 2010).

However, although infants and young children may be able to perceptually discriminate between racial groups, they do not appear to understand the concept of race as signifying a stable and meaningful social category until middle or later childhood (Allport, 1954; Hirschfeld, 1995; Roberts & Gelman, 2016). Moreover, evolutionary arguments suggest that race is a relatively arbitrary stand-in for more relevant cues to social coalitions, and its influence is therefore trumped by competing social categories such as gender, language, or team membership (Kinzler, Shutts, DeJesus, & Spelke, 2009; Kurzban, Tooby, & Cosmides, 2001; Shutts, Banaji, & Spelke, 2010; Sidanius & Pratto, 1999; Weisman, Johnson, & Shutts, 2015).

Additionally, race-specific amygdala activation to other-race faces can dissipate when the faces are familiar (Phelps et al., 2000) or when the perceivers' goals are focused on individuating rather than categorizing (Wheeler & Fiske, 2005). The influence of face-race cues on explicit social judgments can also be overridden by internal or external motivations to respond in socially desirable ways (Devine et al., 2002). Together, these findings of later-emerging conceptual understanding, relative arbitrariness, and the role of motivations and deliberative control together suggest that face-race cues are not mandatory guides in social evaluations.

RELATIVE IMPORTANCE OF FACE-TRAIT VERSUS FACE-RACE CUES

Evidently, competing predictions can be offered on whether face-trait and face-race cues, *in isolation*, will be found to influence explicit social evaluations (i.e., whether they will have independent main effects). However, predictions can also be offered on the *relative* importance of face-trait versus face-race cues (i.e., their interaction), including the theoretical conditions, both experimental and developmental, that give rise to one cue being prioritized over the other.

On the one hand, it is possible that face-trait cues may dominate face-race cues for at least four reasons. First, perceivers make judgments from face-trait cues for faces of other species with which they have limited perceptual familiarity (e.g., rhesus macaques; Cogsdill & Banaji, 2015), implying that face-trait cues may also be sufficiently powerful to override differential familiarity from face-race cues (e.g., the "other race effect"). Second, face-trait cues are argued to reflect evolutionarily adaptive overgeneralizations (Zebrowitz, Bronstad, & Montepare, 2011; Zebrowitz & Montepare, 2008), unlike the evolutionarily arbitrary role of face-race cues (Kurzban et al., 2001). Face-trait cues may therefore be innately and automatically activated.

Third, unlike face-race cues, face-trait cues are relatively unaffected by motivations to correct for prejudice (Devine et al., 2002). Indeed, although race bias is among the most widely discussed prejudices (Charlesworth & Banaji, 2019), "face bias" appears to be much less of a societal concern. Rather, perceivers appear willing to profess judgment on facial physiognomy and show little correction even when given ample time (Willis & Todorov, 2006) or when explicitly confronted with their inaccuracy (Suzuki, 2018; Suzuki, Tsukamoto, & Takahashi, 2019). Thus, while perceivers may correct for the influence of face-race cues to conform to socially desirable responses, it is less likely that perceivers will correct for the influence of face-trait cues. As such, face-trait cues would be expected to be relatively more consequential than face-race cues, particularly in contexts that minimize motivation or opportunity to engage correction processes.

Fourth and finally, beliefs surrounding face-traits appear to be developmentally invariant (Cogsdill & Banaji, 2015; Cogsdill et al., 2014), unlike the tenuous and gradual emergence seen for conceptual racial understanding (Hirschfeld, 1995; Roberts & Gelman, 2016). This would suggest that younger children may be particularly likely to prioritize face-trait cues over face-race cues because of their relatively limited understanding of the social and conceptual implications of race. Additionally, the deliberate use and correction for face-race cues on explicit social judgments increases through middle childhood (Baron & Banaji, 2006; Raabe & Beelmann, 2011), as children become increasingly aware of fairness norms (Rutland, Cameron, Milne, & McGeorge, 2005). Thus, in addition to lower conceptual understanding, younger children may also have lower motivation to use and correct for face-race cues, and therefore may show more consistent prioritization of face-trait cues.

On the other hand, it is possible that face-race cues will appear relatively more important than face-trait cues precisely because of the widespread and pervasive

implicit and explicit race biases (Nosek et al., 2007). That is, despite perceivers' motivations to appear unprejudiced, associations between Black/Afrocentric faces and negative attributes may persist in shaping social judgments, regardless of the face-trait cues evoked by those faces. Indeed, previous research has shown that face-race cues shape the perception of emotion expression, particularly among those with high implicit race bias (Hugenberg & Bodenhausen, 2003), and even among young children (Dunham, 2011; Dunham, Chen, & Banaji, 2013). Face-race cues may therefore dominate the perception of face-trait cues just as they dominate face-emotion cues.

Additionally, face-race cues may be prioritized if face-trait cues are deemed to be particularly irrelevant or non-diagnostic to the judgment, such as when the face-trait cue evokes competence but the judgment is based on warmth or "niceness" (Fiske et al., 2002). Finally, the aforementioned "other race effect" may make it more difficult for perceivers to identify face-trait cues in other-race faces due to differential familiarity, thereby leading face-trait cues to be overridden by face-race cues (Meissner & Brigham, 2001). The present research was designed to examine these competing predictions regarding the prioritization of face-trait and face-race cues in explicit social evaluations.

THE PRESENT RESEARCH

In the current studies, participants viewed face pairs that contrasted cues of face-traits (physiognomic appearance of trustworthiness, submissiveness, or competence) and face-race (physiognomic and skin-tone appearance of Afrocentric or Eurocentric group membership). For each pair, participants chose the face that they judged to be more "nice" (Studies 1–4) or "mean" (Study 3). This design resulted in four types of face pairs: two within-race face pairs consisting of either two White faces (WW) or two Black faces (BB); and two cross-race face pairs of one White and one Black face, in which the Black face evoked the positive traits of trustworthiness, submissiveness, and competence (i.e., Black-Positive/White-Negative, B+W- face pairs) or the Black face evoked the negative traits of untrustworthiness, dominance, and incompetence (i.e., Black-Negative/White-Positive, B-W+ face pairs).

The within-subject factorial design can test theoretically derived patterns that reveal both the independent influence and relative importance of face-trait and face-race cues in guiding social evaluations. Specifically, it is possible to determine whether perceivers (1) always focus on face-trait cues, regardless of face-race cues (i.e., *face-trait cue dominance*); (2) always focus on face-race cues, regardless of face-trait cues (i.e., *face-race cue dominance*); or (3) integrate both face-trait and face-race cues. In the latter pattern, the design can also reveal whether the integration of face-trait and face-race cues indicates race bias (when Black faces are less likely to be picked as "nice") or race correction (when Black faces are more likely to be picked as "nice").

Because multiple patterns can be detected, it is important to investigate the conditions—both experimental and developmental—that characterize when and how perceivers use and integrate face-trait and face-race cues. As such, the present research also evaluates how perceivers' judgments are affected by the motivation and opportunity to engage in correction for possible race bias (Studies 2 and 3), thereby illuminating the possible social psychological mechanisms that underlie the prioritization of such fundamental cues. The research also provides the first examination of how the integration of these cues emerges across early and middle childhood (Study 4), offering new insights into the patterns of change in children's explicit social evaluations.

STUDY 1

Study 1 investigated whether, and if so how, face-trait and face-race cues independently and/or interactively influence adults' evaluative judgments of novel social targets.

METHOD

All measures, manipulations, and exclusions are reported for all experiments; all data were collected in accordance with the university institutional review board protocol. Additionally, all stimuli, data, and analysis scripts are publicly available on the Open Science Framework: https://osf.io/xdt7y/

Participants. Volunteer participants completed the experiment online through the Harvard Digital Lab for the Social Sciences (DLABSS, http://dlabss.harvard.edu). DLABSS participants are volunteers who are primarily recruited through their own online searches as well as through word of mouth, social media platforms, and advertisements. The experiment was listed as a study on "Judgments of Faces," and no deception was used. Participants were entered into a monthly raffle for a $50 Amazon gift card.

The target sample size was 300 participants, a large initial sample to allow for noncompletion and provide sufficient power to detect a small effect. A total of 284 participants began the study online at DLABSS, 85 participants stopped the study before completing all face pairs, and 6 participants did not indicate that they were over 18 years of age, so their data were excluded. These rates of approximately 30% participant noncompletion are comparable to other online volunteer samples (Musch & Reips, 2000; Zhou & Fishbach, 2016).

A final sample of 193 participants was obtained (N_{female} = 102, N_{male} = 74, $N_{other\ gender}$ = 17; M_{age} = 48.5 years, SD_{age} = 16.7 years, range = 20–81 years). The majority of participants identified as White (N = 159), with the remainder identifying as Black/African-American (N = 10), Asian (N = 7), or other race (N = 17). Most participants identified as Liberal (N = 136), although less than half identified as Democrat (N = 84), with the remainder identifying as Republican (N = 30) or Other

(N = 44). The majority had at least a B.A. in their education (N = 135) and were citizens of the U.S. or Canada (N = 166).

A sensitivity power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) indicated that the final sample size had sufficient power (.80) to detect significant odds ratios (OR) of 0.69 and 1.45, or a critical z = +/- 1.64 at an alpha of .05, indicating that the sample was large enough to detect significant small effects with the analytic strategy used.

MATERIALS AND PROCEDURES

Stimuli. Four stimulus sets were generated in FaceGen Modeler using four randomly generated male Black/Afrocentric base faces with Afrocentric physiognomy and skin tone. These four base faces were subsequently transformed to have more prototypical White/Eurocentric physiognomy and skin tone. Notably, transforming from Black appearance to White appearance is equivalent to transforming from White appearance to Black appearance because the same features (i.e., skin tone and physiognomy) would be manipulated in each case to make the same race-prototypical base face. Finally, the base face for each race was transformed to be 3 *SDs* above or below the base face on the appearance of three traits: trustworthiness/untrustworthiness, submissiveness/dominance, and competence/incompetence. These face-trait transformations have been validated with data-driven computational models of adults' judgments to give the clearest signals of their respective traits (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013; Todorov et al., 2008; Todorov & Oosterhof, 2011).

Each stimulus set consisted of 12 individual faces from a 2 (face-race: White/Black) x 3 (face-trait: trustworthiness/submissiveness/competence) x 2 (high or low extremity on trait, e.g., *t*rustworthy/untrustworthy) design. These individual faces were used to generate face pairs, consisting of two faces with the same base face but with one face high and one face low on a given trait. In cross-race face pairs, the two faces also differed on race. Thus, the four possible face pairs were: "White-White" (WW); "Black-Black" (BB); "Black-Positive White-Negative" (B+W-), in which the Black face was facially "nice" (i.e., trustworthy, submissive, or competent); and "Black-Negative White-Positive" (B-W+), in which the White face was facially "nice" (Figure 1). Forty-eight face pairs were therefore generated from a 4 (base face stimulus sets) x 3 (trait: trustworthiness, submissiveness, competence) x 4 (face pair types: WW, BB, B-W+, B+W-) design. The "expected" response was operationalized as picking the facially trustworthy, submissive, or competent face as "nice," regardless of the face-race.

Procedure. Participants viewed all 48 possible face pairs across the four stimulus sets. A random order of face pair presentation was determined by a random number generator, and the order was then reversed to provide two prespecified orders counterbalanced across participants. Before viewing the face pairs, participants were told to select the face they thought was "nice." The prompt "Which of these people is nice?" was written below each face pair for the duration of the study. Following the face selection task, participants responded to two explicit 7-point

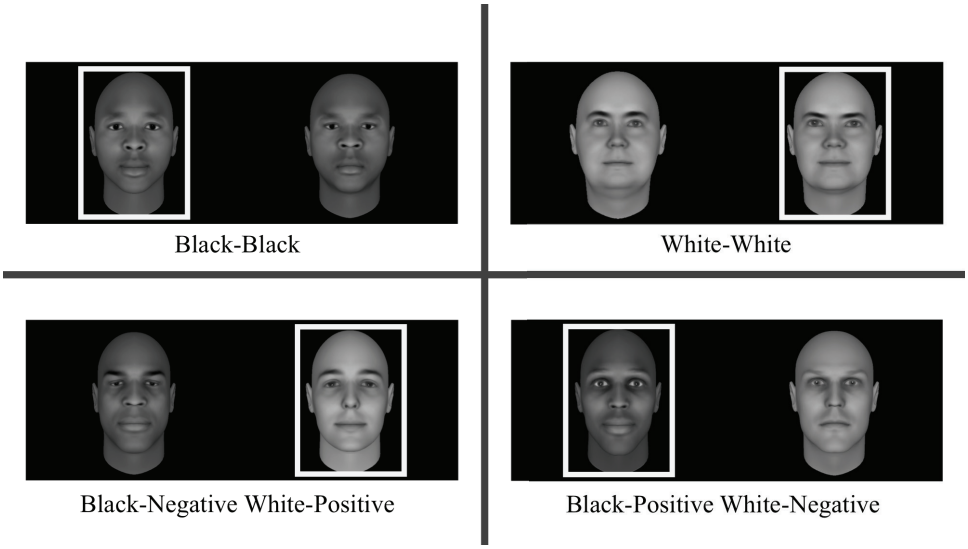

FIGURE 1. Example face pair types presented to participants. An "expected" response for niceness judgments (i.e., selecting the trustworthy, competent, or submissive face as "nice," regardless of face-race) is circled in white.

Likert scales assessing the extent to which they paid attention to (1) facial features and (2) skin-tone/racial category membership in deciding which person was nice, measured from 1 (Not at all) to 7 (A lot). Finally, an exploratory free-response question probed which parts of the face participants attended to. Free responses are not analyzed below but are provided in the open data at https://osf.io/xdt7y/

RESULTS

Analytic Strategy. Overall effects and race-group differences in the percentage selecting the expected (facially nice) face, aggregating at the participant level, are reported below. The influence of face-trait and face-race cues on the likelihood of selecting the expected face was then investigated using a generalized linear mixed-effects model (GLMM) using the *lme4* package with the "bobyqa" optimizer (Bates, Mächler, Bolker, & Walker, 2015) implemented in R (R Core Team, 2017). The binary dependent variable was whether participants chose the "expected" or "unexpected" face based on face-trait cues, as predicted from face-pair type (WW, BB, B-W+, or B+W-). GLMMs are preferable to more traditional ANOVA models primarily because they can incorporate individual trial-level data and do not aggregate at the participant level. Including trial-level data is crucial for experimental designs that could be affected by subtle variations in stimuli (e.g., in this case, stimuli attractiveness, perceived emotion) because trial-level data can control for the specific stimulus face pair used on each trial.

For illustration, if a few face pairs consistently showed high rates of expected responding (perhaps because the face-trait manipulation was particularly visible between those face pairs), then aggregating at the participant-level and collapsing across all stimulus sets could lead to a spurious result driven by only a few face

pairs. However, by maintaining individual trial-level data, the differences in expected responses across face pairs can be included as a random intercept, allowing the model to estimate the effects of interest above and beyond any baseline differences in stimuli. In addition to controlling for such baseline stimulus differences, GLMM random effects can also account for baseline differences across face-traits (trustworthiness, dominance, competence) and across presentation orders.

Given these advantages, the variation across subjects, stimulus sets, face-traits, and order were modeled in a GLMM using random intercepts. First, a random intercept of subject was entered to account for trial-level dependencies from repeated trials provided by the same participant. Second, a random intercept of face pair stimuli (i.e., the specific face pair of the 48 possible pairs used) nested within face-trait (i.e., whether that face varied on competence, trustworthiness, or dominance) was entered to account for possible variation across stimuli, as described above. Third, a random intercept of order (one of two counterbalanced orders, as described in the procedures above) was entered to account for possible baseline differences across orders. The random intercepts were entered successively, and the models were compared using likelihood ratio tests to assess whether the addition of each parameter resulted in a significantly better fit to the data.

Following random effects specification, the fixed effects of interest were entered, as well as covariates of participant gender (male, female, or other) and race (White, Asian, African-American, Other). Covariates were grand-mean centered such that the intercept is interpreted as reflecting the "average" gender or race (Kreft & de Leeuw, 1998), reducing concerns about small sample sizes for estimating the effect of a given racial subgroup. Again, successive fixed effects models were compared using likelihood ratio tests to select the best-fitting and most parsimonious final model. The results of all model comparisons are provided in Supplementary Materials, and the final best-fitting model is described below. The predicted percentages from the final GLMMs are also reported in Table 1 for all studies.

In addition to the main models, we performed supplementary analyses to examine whether the effects were consistent across all face-traits (i.e., trustworthiness, competence, or dominance). Random effects of subject, stimulus pair, and order were entered successively, followed by fixed effects of face pair type (WW, BB, B+W-, B-W+) and face-trait, as well as their interaction. If the observed results are moderated by face-trait, then significant interactions would be expected between face pair and face-trait, and predicted percentages would be expected to indicate different patterns of results as a function of whether the faces varied in competence, trustworthiness, or dominance. The results of the supplementary analyses are provided in detail in the Supplementary Materials and are briefly summarized in the results for each study below.

Overall Effects. Across all face pairs, participants were significantly more likely than chance to select the expected (facially nice) face (M = 78.39%, SD = 10.21%), $t(192) = 38.64$, $p < .001$, $d = 2.78$. There were no differences between White and non-White (Asian, Black, or Other race) participants in overall rates of selecting the expected face (M_{white} = 79.05%, $M_{non\text{-}white}$ = 75.31%), $t(43.26) = 1.73$, $p = .09$, $d = 0.37$. On explicit self-report measures, participants indicated that they rarely used

TABLE 1. Model-Predicted Percentages of Picking the "Expected" Face Based on Face-Trait Cues

	Face Pair							
	White-White		Black-Black		Black-Positive White-Negative		Black-Negative White-Positive	
	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI
Study 1 (adults)	89.33	[81.10%, 94.23%]	89.33	[81.10%, 94.23%]	94.91	[90.46%, 97.34%]	58.44	[42.03%, 73.18%]
Study 2 (adults – self)	89.19	[81.99%, 93.73%]	87.99	[80.18%, 92.99%]	93.66	[89.03%, 96.40%]	63.32	[48.87%, 75.71%]
Study 2 (adults – other)	90.19	[83.52%, 94.33%]	87.86	[79.99%, 92.91%]	77.67	[65.81%, 89.44%]	82.40	[72.13%, 86.28%]
Study 3 (adults – time pressure)	83.59	[75.41%, 89.43%]	83.77	[75.65%, 89.56%]	92.08	[87.42%, 95.10%]	67.70	[55.84%, 77.65%]
Study 4 (kids – youngest age)	81.74	[73.13%, 88.04%]	76.83	[66.94%, 84.44%]	75.91	[64.57%, 83.82%]	74.83	[65.70%, 82.90%]
Study 4 (kids – oldest age)	88.76	[82.59%, 92.94%]	85.72	[78.37%, 90.86%]	94.44	[90.98%, 96.63%]	65.02	[53.14%, 75.29%]

Note. Results are the predictions from the best-fitting models from the generalized linear mixed effects models as outlined in the Results sections for all studies.

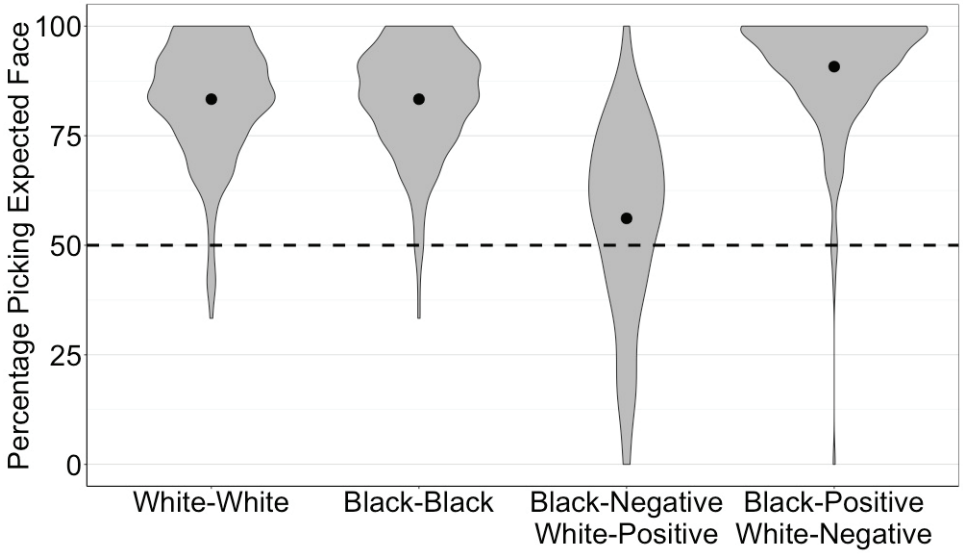

FIGURE 2. The use of face-trait and face-race cues across face pairs. Percentage "expected" indicates the average percentage of trials in which participants selected the trustworthy, competent, or submissive face as "nice." Shaded gray areas reveal the distribution of the raw data; black dots indicate the descriptive mean, collapsing at the participant level.

race information in their niceness judgments (M = 2.21, SD = 1.48), but used face information significantly more frequently (M = 6.22, SD = 1.35), $t(378.92) = -27.79$, $p < .001$, $d = 2.84$.

Effect of Face Pair in Selecting the Expected Face. The model with two random effects of subject and stimuli nested in face-traits significantly improved the model above the previous model with only a random effect for subject, $\chi^2(2) = 1071.05$, $p < .001$. No other random effects offered improvement and were therefore omitted to ensure parsimony and convergence (see Supplementary Materials). Additionally, the model with a fixed effect of face pair (and covariates of gender and race) significantly improved model fit above the covariates-only model, $\chi^2(3) = 1109.7$, $p < .001$.

The final model intercept indicated that participants were significantly more likely than chance to select the expected face in White-White (WW) face pairs, OR = 8.37, $z = 6.23$, $p < .001$ (Table 1, Figure 2). Compared to WW face pairs, participants were equally likely to select the expected face in Black-Black (BB) face pairs, OR = 1.00, $z = 0.00$, $p > .99$. This result newly demonstrates that, contrary to predictions suggesting *face-race cue dominance*, adults appeared to use face-trait cues in social judgments to an equal extent regardless of whether the two faces were White or Black.

However, compared to WW pairs, participants were significantly less likely to select the expected face for the Black-Negative White-Positive (B-W+) face pairs, OR = 0.17, $z = -21.95$, $p < .001$. Recall that, for these face pairs, responses were "expected" if participants selected the White positive face as "nice." Thus, this result suggests that adult perceivers were integrating both face-trait and face-race cues

in their social judgments and were doing so in line with *correction* for possible race bias. That is, in B-W+ face pairs, although participants remained above-chance in selecting the expected face based on face-traits, participants nevertheless showed a reticence to select the positive White face as "nice" when the comparison face was Black, perhaps because such a judgment may be interpreted as calling the Black face "*not* nice" and therefore perceived as race bias.

Analogously, compared to WW face pairs, participants were significantly more likely to pick the expected face in Black-Positive White-Negative (B+W-) pairs, OR = 2.23, $z = 8.15$, $p < .001$. In this case, responses were marked as "expected" if participants selected the positive Black face as "nice." Thus, the high odds indicate that participants were even more likely to select a Black positive face than to select a White positive face. Again, this result implies correction for face-race cues.

Notably, supplemental analyses examining the consistency of these effects across face-traits revealed that the findings were generally consistent across all face-traits: Participants selected the expected face based on face-trait cues and also corrected for face-race cues. Nevertheless, on competence trials participants were least likely to use face-trait cues, and most likely to use and correct for face-race cues. Indeed, on one occasion (B-W+ trials for competence face pairs), participants were below-chance in selecting the expected face based on face-trait cues. Detailed results are reported in Supplementary Materials. This suggests that participants may be attentive to the relevance of the face-trait to their social judgments, a point we address below.

DISCUSSION

The results of Study 1 revealed, first, that adults were consistently above-chance in using face-trait cues and judged the facially nice face (i.e., the trustworthy, submissive, and competent-looking face) as "nice," even when the faces were from a less familiar racial group (i.e., Black/African Americans). Indeed, the results newly show that participants had identical percentages in selecting the expected face based on face-trait cues, regardless of whether the two faces were White or Black. These results replicate prior work on the pervasiveness of face-trait cues (Todorov et al., 2015) and extend beyond such findings to show that face-trait cues are even powerful enough to override differing familiarity with, or stereotypic beliefs about, targets from other racial groups. Studies 2–4 test the robustness of this finding by probing whether the pervasiveness of face-trait cues in Black-Black face pairs is replicated with participants of different ages and with modified experimental designs.

Notably, despite this potency of face-trait cues, Study 1 also indicated that perceivers attend to racial group membership when faces differ on race. That is, face-trait and face-race cues interact in shaping social evaluations, with face-race cues operating primarily through correction for race bias (Devine et al., 2002). Indeed, although participants were consistently above-chance in selecting the face with positive face-traits, they nevertheless displayed a relative reticence to pick the

White face as nice, even when that face evoked positive face-traits. Thus, the second aim of the following studies is to identify the experimental conditions (Studies 2 and 3) and developmental trajectories (Study 4) that characterize such correction processes.

STUDY 2

In Study 1, social evaluations were guided predominantly by face-trait cues. However, face-race cues nonetheless persisted by initiating processes of correction for potential race bias. Study 2 examined whether the influence of face-race cues would be eliminated if participants were less motivated to engage these corrective processes because of reduced concerns about being labeled personally prejudiced. Concerns for personal race bias were manipulated through indirect assessment, in which participants responded from the indirect perspective of an "other," as well as from the direct perspective of the "self." Indirect responding reduces concern of personally expressing prejudice and thus reduces the motivation to control or correct for race bias (Campbell, 1950). If the psychological mechanism behind Study 1 is, indeed, correction for face-race cues, then indirect responding should correspondingly sway participants towards face-trait cues (i.e., *face-trait cue dominance*), and away from face-race cue *correction*.

METHOD

Participants. A total of 375 volunteer participants began the study at DLABSS. A larger target sample than in Study 1 was set to allow for noncompletion and detection of significant interaction effects (specifically, a target of approximately 300 completed participants). Fifty participants left the study before completing all face pairs (87% completion), and the data of 30 participants were omitted for failing to indicate that they were over 18 years of age. Ultimately, a final sample of 295 participants was obtained (N_{female} = 158, N_{male} = 129, $N_{\text{other gender}}$ = 8; M_{age} = 55.6 years, SD_{age} = 16.7 years, range = 20–89 years). The majority of participants identified as White (N = 262), while the remainder identified as Black/African-American (N = 9), Asian (N = 4), or other race (N = 20). Approximately half of the sample identified as Liberal (N = 160), although less than half identified as Democrat (N = 100), with the remainder identifying as Republican (N = 69) or Other (N = 85). The majority also had at least a B.A. in their education (N = 208) and were citizens of the U.S. or Canada (N = 262).

As before, a sensitivity power analysis using G*Power indicated that the final sample size had sufficient power (.80) to detect significant odds ratios of 0.74 and 1.34, or a critical z = +/- 1.64 at an alpha of .05, indicating the sample was large enough to detect significant small effects in a logistic regression.

MATERIALS AND PROCEDURES

Procedures. As in Study 1, participants completed the study online at Harvard DLABSS. Participants viewed all 48 possible face pairs twice: once to respond from the direct perspective of the "self" (i.e., "Which of these people do *you* think is nice?"), and once to respond from the indirect perspective of the "other" (i.e., "Which of these people would *everybody else* think is nice?"). Participants were randomly assigned to see either (1) the "self" and "other" prompts immediately following each other for each face pair ("paired order"); or (2) all "self" prompts followed by all "other" prompts ("blocked order"). Again, the overall presentation of face pairs was determined by a random number generator and then reversed to yield two face pair orders counterbalanced across participants. Thus, four possible orders were used: paired order 1, paired order 2 (the reverse of order 1), blocked order 1, and blocked order 2. As in Study 1, the prompt (self or other) was written below each face pair for each trial. Following the face selection task, participants completed the same explicit measures from Study 1.

RESULTS

Analytic Strategy. A generalized linear mixed-effects model (GLMM) was used to assess the effect of face pair and the interaction with prompt type (self vs. other) on participants' character evaluations. Random effects were specified as in Study 1 with "order" now indicating one of four possible orders (paired or blocked for each of two counterbalanced orders). Fixed effects were prompt (self vs. other) and face pair, their interaction, and grand-mean centered covariates of participant gender and race. Detailed results of all model comparisons are provided in Supplementary Materials, and all predicted percentages from the best-fitting model are presented in Table 1.

Overall Effects. Overall, participants were significantly more likely than chance to select the expected face (M = 79.39%, SD = 8.72%), $t(294) = 57.91$, $p < .001$, $d = 3.37$. There was also an effect of participant race, such that non-White (Asian, Black, or Other race) participants were less likely overall to select the expected face than White participants: M_{white} = 79.87%, $M_{non\text{-}white}$ = 75.60%, $t(37.62) = 2.30$, $p = .03$, d = 0.49, although the actual percentage difference was only 4 percentage points. As in Study 1, participants explicitly indicated that they rarely used race information in their niceness judgments (M = 2.38 SD = 1.53) but used face information significantly more frequently (M = 6.23, SD = 1.21), $t(555.59) = -33.69$, $p < .001$, $d = 2.78$.

Interaction of Face Pair and Prompt Type in Selecting the Expected Face. The model with two random effects of subject and stimulus set nested in trait significantly improved model fit above the previous model with only a random effect of subject, $\chi^2(2) = 3265.02$, $p < .001$. No other random intercepts improved model fit, all $\chi^2 s < 0.003$, all ps > 0.96. The two-way interaction model between face pair type and prompt provided significantly better fit than the previous model with only two main effects, $\chi^2(3) = 724.20$, $p < .001$.

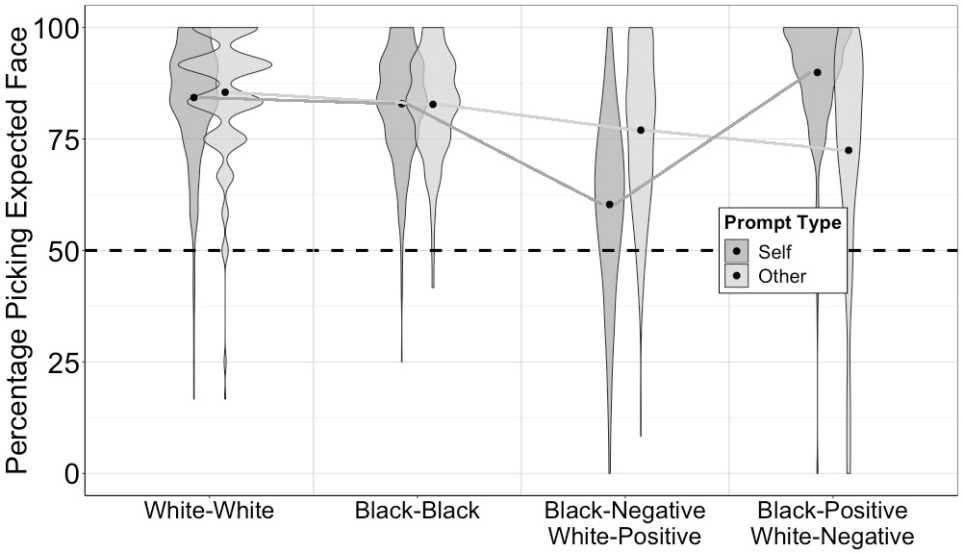

FIGURE 3. The use of face-trait and face-race cues under direct and indirect responding. Percentage picking expected face indicates the average percentage of trials in which participants selected the trustworthy, submissive, or competent face as "nice." Prompt type of "self" was asked from the direct first-person perspective (i.e., "Which of these faces do you think is nice?"); prompt type of "other" was asked from the indirect third-person perspective (i.e., "Which of these faces would everybody else think is nice?"). Shaded gray areas indicate raw distributions of responses; black dots indicate descriptive means, collapsing at the participant level.

Compared to the dummy-coded baseline of White-White face pairs in the "self" prompt, no significant interactions emerged for Black-Black (BB) face pairs, OR = 0.89, z = -1.21, p = .23 (Figure 3), implying that the prompt (self or other) did not alter the use of face-race cues or their corrective processes when the face pairs were within-race.

In contrast, significant interactions emerged between face pair and prompt type for WW versus Black-Negative White-Positive (B-W+) face pairs, OR = 2.43, z = 9.60, p < .001. This shows that, when participants responded from the indirect "other" perspective, the odds of selecting the positive White face were 2.40 times greater in the "other" versus "self" prompt. In other words, for B-W+ face pairs, responding from the "other" perspective reduced correction for face-race cues.

Additionally, significant interactions emerged for WW versus Black-Positive White-Negative (B+W-) face pairs, OR = 0.21, z = -15.25, p < .001. In B+W- face pairs, lowered correction for face-race cues manifests as a *decreased* percentage picking the expected face because participants no longer have the dual motivations to pick the Black face as "nice" to both control prejudice and attend to face-trait cues. Indeed, the odds ratios show that the odds of participants picking the positive Black face as "nice" were 4.76 times greater (i.e., 1 divided by the odds ratio of 0.21) in the "self" versus "other" prompt. Again, this implies that participants were particularly likely to correct for face-race cues (and therefore pick the Black face) when they were concerned with appearing personally prejudiced in the "self" condition.

As in Study 1, supplemental analyses again indicated that the key results were consistent across face-traits of trustworthiness, competence, and dominance (see

Supplemental Materials). Regardless of face-trait, participants in the "self" prompt showed correction for face-race cues, whereas participants in the "other" prompt showed decreased correction and more consistent focus on face-trait cues.

DISCUSSION

The results of Study 2 reinforce the pervasiveness of face-trait cues in guiding social inferences. As in Study 1, participants selected the expected face above-chance at similar rates for both White-White and Black-Black face pairs, regardless of whether they were asked from the direct "self" perspective or the indirect "other" perspective. Thus, the pervasiveness of face-trait cues, even for faces of less-familiar racial groups, is robust to variations in question type and task framing.

The results also revealed that the use, and correction for, face-race cues is reduced when participants have less concern about appearing personally prejudiced. Indirect "other" responses (vs. direct "self" responses) made participants more likely to attend to face-trait cues over face-race cues, suggesting that reducing motivations to correct for possible race bias correspondingly reduced the influence of face-race cues. Notably, although the influence of face-race cues was reduced, the cues nevertheless had a small persistent effect. Specifically, the likelihood to select the expected face remained lower for B-W+ face pairs than for WW face pairs, implying that face-race still influenced adults' evaluative judgments. Foundational work on corrective processes in social inferences (e.g., Gilbert, Pelham, & Krull, 1988) argues that correction is laborious and can only occur with sufficient time and attention. Study 3 therefore examined whether the additional constraint of time pressure will be sufficient to erase correction for face-race cues or whether face-race is sufficiently automatized and powerful to continue to interact with face-traits even in rapid social evaluations.

STUDY 3

Study 3 examined the limits of correction for face-race cues in adults' evaluative judgments by placing participants under time constraints, thereby reducing the opportunity to engage in deliberate correction for face-race cues (e.g., Gilbert et al., 1988). If controlled correction for face-race cues overrides the more automatic use of face-trait cues, then time constraints will sway participants towards face-trait cues alone (i.e., towards *face-trait cue dominance*). If, however, face-race cues remain persistent, perhaps because motivations against prejudice have already been deeply internalized (Devine et al., 2002), then even time constraints may not be sufficient to erase the influence of face-race on evaluative judgments. Study 3 also addressed methodological limitations of previous studies by fully randomizing order of face pair presentation across participants, and by exploring robustness across both "meanness" and "niceness" judgments.

METHOD

Participants. As in Study 1, a target of approximately 300 participants was determined to ensure similar power following data exclusions. A total of 273 participants began the study, 90 participants left the study before completing all face pairs (67% completion), and all participants indicated they were over 18 years of age. A final sample of 183 participants was obtained, (N_{female} = 105, N_{male} = 59, $N_{other\ gender}$ = 19; M_{age} = 50 years, SD_{age} = 18 years, range = 20–90 years). The majority of participants identified as White (N = 148), while the remainder identified as Black/African-American (N = 5), Asian (N = 6), or other race (N = 24). Approximately half of the sample identified as Liberal (N = 94), although less than half identified as Democrat (N = 77), with the remainder identifying as Republican (N = 35) or Other (N = 45). The majority also had at least a B.A. in their education (N = 130) and were citizens of the U.S. (N = 158).

A sensitivity power analysis using G*Power indicated that the final sample size again had sufficient power (.80) to detect significant odds ratios of 0.68 and 1.46, or a critical z = +/- 1.64 at an alpha of .05, indicating that the sample was large enough to detect significant small effects given the analytic strategy used.

Materials and Procedures. All participants completed the study online at DLABSS. Participants viewed all 48 possible face pairs twice: once to evaluate "niceness," and once to evaluate "meanness." Participants rated "niceness" and "meanness" in the same single block of trials with the order of the face pairs randomized across participants, including whether a given face pair was first evaluated on "niceness" or "meanness." The prompt ("nice" or "mean") was written above each face pair in large letters. Additionally, participants were told that they would be required to respond within 2500 ms, a time limit chosen to ensure that participants were under time pressure to read the prompt (nice or mean), move the cursor, and select the face, but were still able to respond within the time limits for the majority of trials. If participants did not respond within this time window, the page would flash an error to "Go Faster." Latencies were recorded for both the first click and the final page submission.

Because of task difficulty, participants were given 10 practice trials (5 nice, 5 mean), with novel White-White face pairs that did not appear in the main task. Data from these practice trials were not included in any analyses. Following the main face selection task, participants responded to four explicit questions assessing the degree to which they used face or race information when making niceness and meanness judgments, from 1 (Not at all) to 7 (A lot).

RESULTS

Analytic Strategy. Only trials with latencies less than or equal to 2500 ms were included in the analysis, ensuring that all analyzed responses conformed to the requirement of the time constraints and rapid responding. This led to an exclusion of

approximately 24% of trials overall. Importantly, supplemental analyses including *all* trials provide the same conclusions as the analyses presented below using only rapid trials. This reduces concerns that participant attrition or trial elimination could account for the observed results (see Supplementary Materials).

As before, a generalized linear mixed-effects model (GLMM) was used. Random intercepts were entered as in Study 1, with two exceptions: (1) no random intercept of order was included since order was fully randomized across participants; and (2) a random intercept of judgment type ("nice" or "mean") was added. The fixed effects for Study 3 were face pair type as well as grand-mean centered covariates of gender and race. Detailed results of all model comparisons are provided in Supplementary Materials and predicted percentages from the final model are reported in Table 1.

Overall Effects. Across all face pair types, participants were significantly more likely than chance to select the expected face, even when required to respond within 2500 ms (M = 78.69%, SD = 7.90%), $t(182) = 49.09$, $p < .001$, $d = 3.63$. There were no differences between White and non-White (Asian, Black, or Other race) participants in overall rates of selecting the expected face: M_{white} = 78.89%, $M_{non\text{-}white}$ = 78.57%, $t(48.68) = 0.22$, $p = .83$, $d = 0.04$.

As in Studies 1 and 2, participants explicitly indicated that they used race information (M = 2.19, SD = 1.53) significantly less frequently than face information (M = 6.42, SD = 1.07) in their niceness judgments, $t(314.71) = -30.22$, $p < .001$, $d = 3.21$. An identical pattern was observed for meanness judgments: Participants indicated that they used race information (M = 2.06, SD = 1.41) significantly less frequently than face information (M = 6.44, SD = 1.13), $t(338.95) = -32.37$, $p < .001$, $d = 3.43$.

Effect of Face Pair in Selecting the Expected Face Under Time Constraints. The model with three random intercepts of subject, stimulus set nested within trait type, and judgment type provided significant improvement over the previous model with two random intercepts of subject and stimuli, $\chi^2(1) = 8.42$, $p = .004$. Additionally, the fixed effect of face pair type provided significantly better fit than the covariates-only model, $\chi^2(3) = 648.82$, $p < .001$.

The final model indicated that, for rapid (< 2500 ms) evaluations, participants were equally likely to pick the expected face in both White-White and Black-Black face pairs, OR = 1.02, $z = 0.30$, $p = .77$. The pervasiveness of face-trait cues for less-familiar racial groups therefore appears robust even to time constraints and variations in experimental design.

Compared to judgments in WW face pairs, participants also remained less likely to select the expected face in Black-Negative White-Positive face pairs, OR = 0.41, $z = -14.46$, $p < .001$, as well as more likely to select the expected face in Black-Positive White-Negative face pairs, OR = 2.28, $z = 10.80$, $p < .001$. Thus, a pattern of correction for face-race cues persisted even when participants were placed under time constraints. Moreover, supplemental analyses showed that these results were again largely consistent across all face-traits (see Supplementary Materials).

DISCUSSION

Study 3 further demonstrated the pervasiveness of face-trait cues in guiding social judgments across both White and Black faces. Even within 2500 ms, participants were significantly more likely than chance to select the expected face based on face-trait cues as "nice" or "mean," regardless of the faces' racial appearance. Moreover, Study 3 demonstrated that time constraints had a small effect on disrupting the processes of correction for face-race cues, slightly increasing participants' tendency to select the "nice" or "mean" face based on face-trait cues rather than face-race cues in Black-Negative White-Positive face pairs.

Nevertheless, as with indirect responding in Study 2, time constraints did not fully eliminate the persistent use of face-race cues in cross-race face pairs. Indeed, in cross-race face pairs, participants remained more likely to pick the Black face as "nice" and less likely to pick the White face as "nice." Previous work has similarly found that time constraints do not alter the consistency of judgments from face-based cues: Judgments made within a 2000 ms time window were not substantially different from those made without a response deadline (Ballew & Todorov, 2007). This suggests that face-trait cues may be used rapidly and without deliberation. We newly show that face-*race* cues may also be rapidly used and corrected for, perhaps because participants have deeply internalized motivations to appear unprejudiced (Devine et al., 2002).

Additionally, given that above-chance consistency in face-trait inferences has been observed even when faces are presented for only 33 ms (Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006), it is possible that 2500 ms provided ample time for participants to make an evaluative judgment from face-trait cues, notice racial group differences, and still engage in subsequent deliberative correction for face-race cues. Future research using masking procedures and more rapid presentation may be able to identify the boundary at which the correction for face-race cues no longer influences adults' evaluative judgments.

STUDY 4

Thus far, three studies have demonstrated that face-trait cues guide social evaluations, with participants consistently selecting the expected face based on face-trait cues, even when the faces were from a less-familiar racial group. This provides support for the relative importance of face-trait cues, even above the widely discussed face-race cues. Nevertheless, these studies also find that face-race cues exert surprisingly persistent influences through corrective processes, suggesting that adults may rapidly encode, use, and control for the effect of face-race cues on their explicit evaluations.

Such rapid encoding of, and correction for, face-race cues may reflect prolonged social learning about race relations and prohibitions against the use of race in

explicit social judgments. Indeed, developmental research shows that an understanding of the social implications of race emerges only gradually through middle childhood (Allport, 1954; Fitzroy & Rutland, 2010; Hirschfeld, 1995; Raabe & Beelmann, 2011; Roberts & Gelman, 2016; Rutland et al., 2005; Rutland, Killen, & Abrams, 2010). As such, attention to, and correction for, face-race cues may not appear in young children who have only tenuous understandings of social norms and race as a social concept. Instead, face-trait cues may completely override face-race cues among young children—a pattern of prioritization that was not obtained even after placing adults under time constraints (Study 3) and reduced concerns to appear unbiased (Study 2). A developmental analysis is therefore crucial to examine the possible limits on the influence of face-trait and face-race cues in social evaluations, as well as to identify the emergence of the interaction between these two fundamental cues.

Study 4 examines age-related changes from 5 to 13 years of age in the use of face-trait and face-race cues to illuminate (1) whether face-trait cues powerfully guide social evaluations across development, and (2) whether the persistent influence of, and correction for, face-race cues is eliminated among the youngest children, for whom conceptual racial understanding and social desirability concerns are still emergent.

Given that infants as young as 7 months of age indicate preferences for faces with positive facial traits (Jessen & Grossmann, 2016), it is predicted that children, like adults in Studies 1–3, will show consistent use of face-trait cues in guiding social judgments for all within-race face pairs. With respect to emerging correction for face-race cues, past research has revealed that explicit preference for White Americans decreases through middle childhood, alongside increasing concerns of social desirability and fairness norms (Baron & Banaji, 2006; Raabe & Beelmann, 2011; Rutland et al., 2005) and increasing conceptual racial knowledge (Allport, 1954; Hirschfeld, 1995; Roberts & Gelman, 2016). If the persistent influence of face-race cues in social judgments is driven by concurrent developments in concerns for fairness and conceptual racial understanding, then correction for face-race cues should not be present among the youngest children. Instead, the youngest children should exhibit a pattern of *face-trait cue dominance.*

METHOD

Participants. A total of 311 participants participated at public elementary and middle schools in Victoria, BC, Canada, as well as at public parks in Cambridge, Massachusetts, and a public children's museum in Boston, Massachusetts. A target of 150 child participants was set per location. Notably, the two testing locations differ in the prevalence of their Black populations: Victoria, BC, has a Black/African-Canadian population of approximately 1% (Statistics Canada, 2016), whereas Cambridge has a Black/African-American population of approximately 11% (U.S. Census Bureau, 2016). Although not central to the current research, similarity in children's use of face-trait and face-race across these two locations would help ad-

dress concerns about the robustness of the phenomena across geographic contexts and racial exposure.

Data from 18 child participants were omitted due to: noncompletion (2 participants), parental interference (4), random responding or not following instructions (4), and being outside the pre-specified age range of 5–13 (8). This yielded a total final sample size of 293 participants (N_{female} = 155; $N_{Victoria}$ = 150; M_{age} = 9.35 years, SD_{age} = 2.22 years, range = 5.03–13.67 years). The distribution of participants across ages was as follows: 5–8 years (N = 91), 9–11 years (N = 122), and 12–14 years (N = 80). The majority of participants identified or were identified by their parents as White (N = 205), while the remainder identified or were identified by their parents as Black/African-American or African-Canadian (N = 12), East Asian (N = 56), or other race (N = 20).

A sensitivity power analysis using G*Power indicated that the final sample size had sufficient power (.80) to detect a critical z = +/- 1.64 at an alpha of .05, or significant odds ratios of 0.74 and 1.35 in a logistic regression, indicating that the sample was again sufficient to detect significant small effects.

Materials and Procedures. All stimuli and procedures were identical to Study 1, except that children participated in person with an experimenter sitting on the opposite side of the computer. Additionally, due to time limitations not all child participants were able to complete the explicit questions; explicit data are therefore excluded from child analyses. Available raw responses are provided at https://osf.io/xdt7y/

RESULTS

Analytic Strategy. As in Study 1, a generalized linear mixed-effects model (GLMM) was used. Random effects were specified in the same order as in Study 1 but with the addition of testing country (US or Canada). Fixed effects were face pair and age (continuous from 5 years to 13 years), as well as their interaction and grand-mean centered covariates of participant gender and race. Detailed results of all model comparisons are provided in Supplementary Materials, and the predicted model percentages are reported in Table 1.

Overall Effects. Across all face pair types, children were significantly more likely than chance to select the expected face (M = 77.45%, SD = 11.20%), $t(292) = 41.94$, $p < .001$, d = 2.45. There were no differences between White and non-White (Asian, Black, or Other race) participants in overall rates of selecting the expected face: M_{white} = 78.00%, $M_{non\text{-}white}$ = 76.16%, $t(162.49) = 1.28$, $p = .20$, d = 0.16. There were also no differences by country of respondents (M_{Canada} = 77.00%, M_{USA} = 77.91%), $t(298.05) = -0.70$, $p = .49$, d = 0.082, suggesting that the patterns of children's face-trait and face-race judgments are robust even across slight variations in racial context.

Interaction of Face Pair and Age in Selecting the Expected Face. The model with two random effects of subject and stimulus set nested within face trait significantly improved model fit above the model with a single random effect of subject, $\chi^2(2) =$

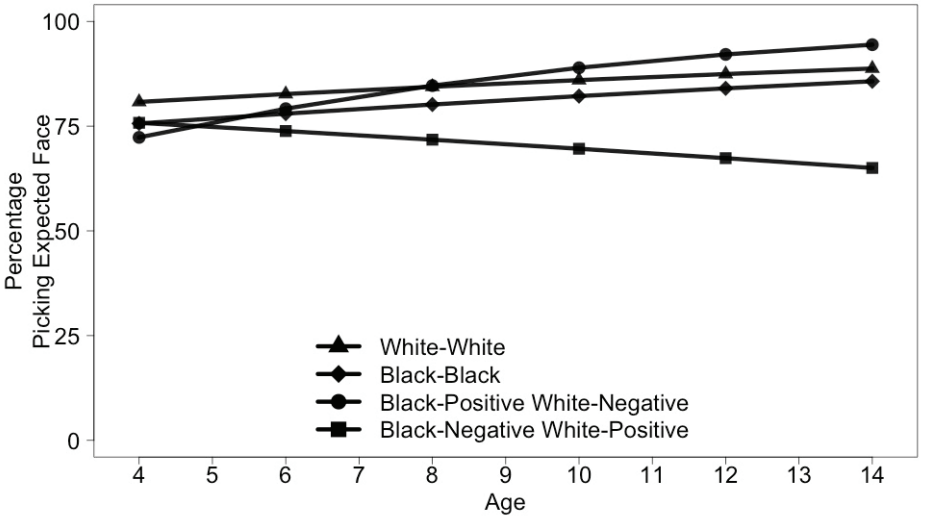

FIGURE 4. Developmental change in the use of face-trait and face-race cues across face pair types. Percentage picking expected face indicates the model-based predicted percentage of trials in which participants selected the trustworthy, submissive, or competent face as "nice," derived from the interaction model with age and face pair type as fixed effects, and subject, stimulus set, and trait type as random intercepts.

1071.8, $p < .001$; no other random intercepts improved model fit, all $\chi^2 s = 0$, all $ps > .99$, within machine precision. The interaction model provided significantly better fit above the previous model with two main effects, $\chi^2(3) = 74.51$, $p < .001$.

Compared to White-White (WW) face pairs and the youngest children, no significant interaction emerged for Black-Black (BB) face pairs, OR = 1.00, $z = 0.10$, $p = .92$ (Figure 4), indicating that the rate of age-related change was parallel in both WW and BB face pairs. This reinforces the early developmental emergence of face-based evaluations (Cogsdill et al., 2014), and newly shows that this early use of face-trait cues is robust across both White and Black faces.

In contrast, significant interactions between face pair and age emerged for Black-Negative White-Positive (B-W+) face pairs, OR = 0.89, $z = -4.27$, $p < .001$, as well as for Black-Positive White-Negative (B+W-) face pairs, OR = 1.13, $z = 4.15$, $p < .001$. Specifically, in B-W+ face pairs, with each year of age the odds that children would select the White positive face *decreased* by 1.12 (i.e., 1 divided by the OR of 0.89). In B+W- face pairs, however, with each year of age the odds that children would select the Black negative face *increased* by 1.13.

Moreover, inspection of the model's predicted values (see Table 1) indicates that children at the youngest ages were selecting the expected face at equal rates across all face pairs, implying a pattern of *face-trait cue dominance.* With age, however, children showed a gradual emergence of the use of, and correction for, face-race cues in social evaluations (Table 1).

Finally, as with all other studies, the key results were consistent across all face-traits. Regardless of whether the faces varied in trustworthiness, dominance, or

competence, children at the youngest ages focused on face-trait cues and showed no evidence of using face-race cues through correction processes. Additionally, across all face-traits, face-race correction emerged gradually through middle childhood.

DISCUSSION

Children from the youngest ages showed consistency across White-White and Black-Black face pairs, highlighting the early emergence and pervasiveness of face-trait cues, even for faces from less-familiar racial groups. Indeed, age-related changes had parallel rates in both White-White and Black-Black face pairs, newly suggesting that developmental changes in the use of face-trait cues occur regardless of whether the faces are White or Black.

In cross-race face pairs, however, children showed (1) age-related *decreases* in the likelihood to select the facially nice White face as "nice" (in B-W+ face pairs); as well as (2) age-related *increases* in the likelihood to select the facially nice Black face as "nice" (in B+W- face pairs). These two trends together indicate that the use and correction for face-race cues emerges progressively across middle childhood.

The timing of age-related changes in control for face-race cues appears concurrently with changes in middle childhood regarding sensitivity to fairness norms (Rutland et al., 2005), control of explicit prejudice (Baron & Banaji, 2006; Raabe & Beelmann, 2011), and conceptual understanding of race as a stable and meaningful social category (Roberts & Gelman, 2016). Thus, in early childhood, children appear to exhibit *face-trait cue dominance,* yet as soon as face-race cues begin to be used, children show an intention to correct for possible bias. Importantly, at no point in development did children show evidence in support of the opposite trends of *face-race cue dominance* or the interactive use of face-trait and face-race cues in line with race bias.

GENERAL DISCUSSION

FACE-TRAIT CUES: EARLY-EMERGING AND PERVASIVE

Facial appearance (i.e., physiognomic cues of perceived trustworthiness, submissiveness, and competence) was found to be a powerful guide in adult's and children's evaluative face-based judgments. The strength of such face-trait cues is seen, first, in the extent to which the cues generalize across judgments of both White/Eurocentric and Black/Afrocentric targets. Adults (Studies 1–3) and children (Study 4) were consistently above-chance in choosing the "nice" face based on face-trait cues in within-race face pairs, regardless of whether they were choosing between two Black or two White faces. These findings were also observed regardless of whether the faces varied on competence, dominance, or trustworthiness. Moreover, even for cross-race face pairs (Black-Positive White-Negative, and Black-Negative White-Positive) in which the additional complexity of contrasting

face-race cues could have swayed participants away from face-trait cues, participants generally selected the "nice" face based on face-trait cues significantly above chance.

The power of face-trait cues across racial groups is also seen in their early emerging use. From the earliest ages tested (5 years old), participants consistently chose the facially "nice" face for all face pairs. Moreover, children showed parallel increases with age in selecting the expected face for both within-race (Black-Black and White-White) face pairs. These findings extend research on children's (Charlesworth, Hudson, Cogsdill, Spelke, & Banaji, 2018; Cogsdill et al., 2014) and infants' (Jessen & Grossmann, 2016) spontaneous facial evaluations by showing that such face-based evaluations are not empty associations or arbitrary responses. Instead, even without extensive social learning, face-trait cues appear to be perceived as sufficiently meaningful to override differential familiarity with, or beliefs about, other racial groups.

The robustness of face-trait inferences across racial groups may appear, at first, inconsistent with a strong interpretation of the "other-race effect" (ORE; Meissner & Brigham, 2001). That is, the ORE would predict that the current majority-White sample would be less likely to select the expected face for Black-Black face pairs because of difficulty disambiguating face-trait cues among faces from less-familiar racial outgroups. Instead, the current results (and others, e.g., Cogsdill & Banaji, 2015) imply that identifying and using latent face-trait cues may be rapid and relatively unconstrained by differential familiarity.

Moreover, we argue that the results actually align with recent social cognitive theories of the ORE that propose that similar face-processing can be achieved for both own- and other-race faces when perceivers have sufficient motivation to individuate the targets (Young, Hugenberg, Bernstein, & Sacco, 2012). The current research contributes further psychological conditions—specifically, the motivation to correct for possible race bias—that may facilitate face-based judgments for less-familiar racial groups.

FACE-RACE CUES: LATER-EMERGING AND CONTROLLED CORRECTION

In comparison to face-trait cues, face-race cues followed a later-emerging trajectory and exerted their influence predominantly through control for racial bias. Participants displayed a relative tendency to pick the Black face as "nice," and a reticence to pick the White face as "nice" for cross-race face pairs. The role of control was highlighted in Studies 2 and 3, where the correction for face-race cues was reduced among adults responding from an indirect "other" perspective (rather than the direct "self," Study 2), as well as among adults placed under time pressure (Study 3). This suggests that reducing adults' personal motivation and opportunity to control prejudice correspondingly reduces the influence of face-race cues on explicit face-based judgments.

The mechanism of control is further suggested by the observed age-related changes in the integration of face-trait and face-race cues (Study 4). The youngest

children (5 years old) did not appear to use face-race cues in their social judgments, showing nearly identical rates of selecting the expected face based on face-trait cues, regardless of the face's racial group. The use of face-race cues emerged gradually through middle childhood and increased into adulthood. This developmental timing aligns with previously documented trajectories of increasing knowledge about race as a stable and meaningful conceptual category (Allport, 1954; Hirschfeld, 1995; Roberts & Gelman, 2016) as well as about fairness norms to control prejudice (Fitzroy & Rutland, 2010; Rutland et al., 2005, 2010). The age-related changes also extend prior work on the development of children's control for prejudice on explicit race attitudes (Baron & Banaji, 2006; Raabe & Beelmann, 2011), race categorization tasks (Apfelbaum, Pauker, Ambady, Sommers, & Norton, 2008), and resource allocations across racial groups (Elenbaas & Killen, 2016). That such age-related change also appears on facial evaluation tasks suggests that the developmental emergence of race-based correction is generalizable and robust across judgment types, even for judgments derived from powerful and early-emerging face-trait cues.

Note that it could have been that perceivers of all ages would use face-race cues *in line* with widespread racial biases (Nosek et al., 2007), leading to a tendency to select the White face as "nice" and Black face as "mean." Indeed, this pattern would be expected given research on the role of face-race and face-trait cues in racial and emotion categorization judgments, where a Black face is more likely to be categorized as "angry" than "happy" and an angry face is more likely to categorized as Black (Dunham et al., 2013; Hugenberg & Bodenhausen, 2003, 2004). However, the current research differs from these studies showing exaggeration of race bias in many ways, including (1) using latent face-traits rather than emotion expressions and (2) evaluative rather than categorization judgments. Indeed, it could be that evaluative judgments, but not categorization judgments, activate corrective processes because the former are explicitly value-laden and may therefore be particularly likely to activate concerns about appearing prejudiced. Future research on the interaction of face-trait and face-race cues would benefit from considering many stages of social judgments, including categorization, evaluation, and correction (Gilbert et al., 1988).

IMPLICATIONS AND LIMITS OF FACE-BASED BIASES

It is worth noting that, despite participants' consistent correction for race bias, there was no indication of correction for "face bias." Although widely used, physiognomic face-trait cues have little to no predictive accuracy of an individual's actual character (Olivola et al., 2014; Todorov et al., 2015), and therefore using face-trait cues as the basis for evaluation presents a source of social bias. Thus, if participants were generally concerned with biased judgments, they could have chosen faces randomly or, as was the case for face-race, chosen the face that was unexpected (i.e., the facially mean face). Yet participants in the present studies (as well as many others, Todorov, 2017; Todorov et al., 2008) consistently used face-

trait cues to guide their evaluations of others, with little variation across ages, question types, or time constraints. Whether right or wrong, the use of face-trait cues remains less of a personal or societal concern than the use of race in evaluations.

While face-based biases are not as frequently discussed as race-based biases, they nevertheless have the potential to cause substantial unfairness and discrimination in real-world outcomes ranging from success in elections (Antonakis & Dalgas, 2009; Lawson, Lenz, Baker, & Myers, 2010; Todorov et al., 2005) to hiring and compensation (Graham, Harvey, & Puria, 2017) to the harshness of extreme criminal sentences (Porter, ten Brinke, & Gustaw, 2010; Wilson & Rule, 2015). Even when all else is equal, more competent or trustworthy-looking targets appear to gain social advantages.

Given these pernicious consequences of face-based biases, it is important to consider whether, and if so under what conditions, face-trait cues may *not* predominate in social evaluations or may even be corrected for. First, it is possible that face-trait cues are less likely to be used, and more likely to be corrected for, when perceivers deem face-trait cues to be irrelevant to the judgment at hand. For example, perceivers may see face-trait cues of competence to be irrelevant to judgments of warmth and "niceness" because competence and warmth are argued to be unrelated (Fiske et al., 2002). Indeed, our supplemental analyses show that participants making niceness judgments were least likely to use face-trait cues (and most likely to use and correct for face-race cues) when the faces varied in competence, suggesting that participants were attentive to the relevance of the face-trait cues in their social judgments.

Relatedly, it is likely that perceivers may negate the influence of face-trait cues when additional available cues are deemed to be particularly diagnostic. For example, if perceivers are given information about a target's past behaviors, such as that they consistently share their money with others (diagnostically indicating trustworthiness), then perceivers may preferentially use this diagnostic information over the relatively more ambiguous information from facial appearance.

Finally, the use of face-trait cues may be reduced if perceivers have increased awareness and concern about the consequences of their own face-based biases. As we argue above, the motivation to correct for face-race cues may arise from the frequent discussion and societal sanctioning of race biases. It is therefore possible that exposing participants to similar discussions on face-based biases, as well as evidence of participants' own biases (e.g., Suzuki, 2018), may correspondingly increase participants' motivation to correct for the influence of face-trait cues and thereby reduce the unfair treatment that may arise from face-based biases.

REMAINING QUESTIONS

This research reveals the developmental trajectory and social psychological mechanisms of evaluative judgments based on face-trait cues across White and Black faces. In so doing, the studies open many avenues for further exploration. First,

the documented correction for face-race cues may be specific to the Black–White comparison, a race relation that is historically and presently contentious in North America, and thus particularly prone to social desirability concerns. Future work exploring the use of face-race and face-trait cues with White versus Asian or White versus Hispanic faces would provide informative comparisons to examine the generalizability across differences in groups' social status and differences in the extent to which it is considered socially undesirable to express bias towards those groups (Dunham, Baron, & Banaji, 2007; Sidanius & Pratto, 1999).

Generalizability of the results could also be explored by including stimuli that differ on age or gender. For instance, it could be that child perceivers are more likely to engage correction processes when judging faces of child peers than when judging faces of adults. Additionally, the current studies used only male stimuli in order to control for documented confounds between gender and perceived latent traits (e.g., female stimuli are perceived as more "trustworthy," Todorov, 2017). Yet future research could benefit from exploring the intersection of race and gender in shaping the use of face-race cues in social evaluations, given findings of "gendered race" (Johnson, Freeman, & Pauker, 2012; Schug, Alt, & Klauer, 2015; Sesko & Biernat, 2010). Indeed, it may be that the observed effects are stronger for male than for female faces because "Black" is gendered as "male" and therefore face-race cues of Black faces may be particularly noticeable (and likely to be corrected for) when the targets are male. Finally, robustness across computer-generated and naturalistic stimuli remains to be explored, although previous studies have suggested little variation between children and adults' face-based inferences across naturalistic, computer-generated, or even other species' facial stimuli (Cogsdill & Banaji, 2015).

Future research may also benefit from examining the role of the perceiver's racial group membership: While few meaningful differences were observed by participant race, the present research was limited by using a predominantly White sample. Focusing on the responses of Black participants could reveal whether correction for race bias is motivated by (1) correction for in-group bias, such that Black participants show a reticence to pick their in-group Black face as "nice," or (2) a broader societal concern with historical inequality, such that Black and White participants show the same processes of correction for face-race cues. Additionally, examining the differences in developmental trajectories across perceiver racial groups might reveal that Black children show earlier-emerging trajectories of correction for face-race cues, given their earlier-emerging conceptual racial understanding in other tasks (e.g., Roberts & Gelman, 2016).

Finally, the current work cannot speak to the individual or developmental correlates of the use and correction for face-trait and face-race cues. Are those with high explicit or implicit preferences for Whites more or less likely to show correction for face-race cues (Hugenberg & Bodenhausen, 2003, 2004)? Is the observed age-related change the result of increasing social norm awareness, racial understanding, or both? Directly exploring these social psychological correlates and concurrent developmental trajectories will help researchers understand the mechanisms of two fundamental cues—face-trait and face-race—in shaping social evaluations.

CONCLUSION

In the early 1800s, the Swiss physiognomist Johann Caspar Lavater remarked that, "whether they are or are not sensible of it, all men are daily influenced by physiognomy" (p. 9, Lavater, 1772); more than 200 years later, research by Todorov (2017), Zebrowitz (2017), Freeman and Johnson (2016), and others has revealed this remark to be true in striking ways. The current experiments continue to provide support by showing that face-trait cues are not only pervasive and early-emerging guides (Cogsdill et al., 2014) but are also relatively stronger than face-race cues in explicit social judgments. Face-trait cues are used from the earliest ages tested and generalize across both Black and White face pairs. Face-race cues, in contrast, are used gradually through middle childhood and adulthood by activating perceivers' correction for possible racial bias. Together, these findings contribute to an understanding of the origins and consequences of explicit social evaluations. The present research reveals both the distinct, and interactive, social psychological mechanisms and age-related changes of face-trait and face-race cues in shaping human's evaluations of others.

REFERENCES

Aboud, F. E. (1988). *Children and prejudice*. Oxford, UK: Blackwell.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*(5918), 1183. https://doi.org/10.1126/science.1167748

Apfelbaum, E. P., Pauker, K., Ambady, N., Sommers, S. R., & Norton, M. I. (2008). Learning (not) to talk about race: When older children underperform in social categorization. *Developmental Psychology*, *44*(5), 1513–1518. https://doi.org/10.1037/a0012835

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948–17953. https://doi.org/10.1073/pnas.0705435104

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and nurture in own-race face processing. *Psychological Science*, *17*(2), 159–163. https://doi.org/10.1111/j.1467-9280.2006.01679.x

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, *17*(1), 53–58. https://doi.org/10.1111/j.1467-9280.2005.01664.x

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

Campbell, D. T. (1950). The Indirect Assessment of Social Attitudes. *Psychological Bulletin*, *47*(1), 15–38. https://doi.org/10.1037/h0054114

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174-192. https://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S., Hudson, S.-K., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2018). Children use targets' facial appearance to guide and predict social behavior. *Submitted for publication*.

Clark, K. B., & Clark, M. P. (1947). Racial identification and preference in Negro children. In E. E. Maccoby, T. M. Newcomb, & E. L. Hartley (Eds.), *Readings*

in social psychology (pp. 602–611). New York, NY: Henry Holt. https://doi.org/10.2307/2966491

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. https://doi.org/10.1016/j.jesp.2015.05.007

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. https://doi.org/10.1177/0956797614523297

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835–848. https://doi.org/10.1037/0022-3514.82.5.835

Dunham, Y. (2011). An angry = Outgroup effect. *Journal of Experimental Social Psychology*, *47*(3), 668–671. https://doi.org/10.1016/j.jesp.2011.01.003

Dunham, Y., Baron, A. S., & Banaji, M. R. (2007). Children and social groups: A developmental analysis of implicit consistency in Hispanic Americans. *Self and Identity*, *6*(2–3), 238–255. https://doi.org/10.1080/15298860601115344

Dunham, Y., Chen, E. E., & Banaji, M. R. (2013). Two signatures of implicit intergroup attitudes. *Psychological Science*, *24*(6), 860–868. https://doi.org/10.1177/0956797612463081

Elenbaas, L., & Killen, M. (2016). Children rectify inequalities for disadvantaged groups. *Developmental Psychology*, *52*(8), 1318–1329. https://doi.org/10.1037/dev0000154

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. https://doi.org/10.3758/BRM.41.4.1149

Fiske, S. T., Cuddy, A. J. C., 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*(6), 878–902. https://doi.org/10.1037//0022-3514.82.6.878

Fitzroy, S., & Rutland, A. (2010). Learning to control ethnic intergroup bias in childhood. *European Journal of Social Psychology*, *40*(4), 679–693. https://doi.org/10.1002/ejsp.746

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences*, *20*(3), 362-374. https://doi.org/10.1016/j.tics.2016.03.003

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740. https://doi.org/10.1037/0022-3514.54.5.733

Glaser, J. (2015). Suspect race: Causes and consequences of racial profiling. New York: Oxford University Press.

Graham, J. R., Harvey, C. R., & Puria, M. (2017). A corporate beauty contest. *Management Science*, *63*(9), 3044–3056. https://doi.org/10.1287/mnsc.2016.2484

Hehman, E., Flake, J. K., & Calanchini, J. (2017). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, 1–9. https://doi.org/10.1177/1948550617711229

Hirschfeld, L. A. (1995). Do children have a theory of race? *Cognition*, *54*(2), 209–252. https://doi.org/10.1016/0010-0277(95)91425-R

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*(6), 640–643. https://doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342–345. https://doi.org/https://doi.org/10.1111/j.0956-7976.2004.00680.x

Jessen, S., & Grossmann, T. (2016). Neural and behavioral evidence for infants' sensitivity to the trustworthiness of faces. *Journal of Cognitive Neuroscience*, *28*(11), 1728–1736. https://doi.org/10.1162/jocn_a_00999

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–131. https://doi.org/10.1037/a0025335

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., … Pascalis, O. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*(6), F31–F36. https://doi.org/10.1111/j.1467-7687.2005.0434a.x

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition*, *27*(4), 623–634. https://doi.org/10.1521/soco.2009.27.4.623

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage. https://doi.org/10.4135/9781849209366

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*(26), 15387–15392. https://doi.org/10.1073/pnas.251541498

Lavater, J. C. (1772). *Essays on physiognomy; for the promotion of the knowledge and the love of mankind*. (T. Holcroft, Ed.). (2nd ed.). London, UK: C. Whittingham for H.D. Symonds.

Lawson, C., Lenz, G. S., Baker, A., & Myers, M. (2010). Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics*, *62*(04), 561–593. https://doi.org/10.1017/s0043887110000195

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Blacks' death rate due to circulatory diseases is positively related to Whites' explicit racial bias. *Psychological Science*, *27*(10), 1299–1311. https://doi.org/10.1177/0956797616658450

Meissner, C. A., & Brigham, J. C. (2001). Own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037//1076-8971.7.1.3

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. *Psychological Experiments on the Internet*, 61–87. https://doi.org/10.1016/B978-012099980-4/50004-6

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology*, *108*(5), 937–975. https://doi.org/10.1086/374403

Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*(2), 367–374. https://doi.org/10.1016/j.jesp.2009.11.001

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. https://doi.org/10.1162/089892900562552

Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime and Law*, *16*(6), 477–491. https://doi.org/10.1080/10683160902926141

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Raabe, T., & Beelmann, A. (2011). Development of ethnic, racial, and national prejudice in childhood and adolescence: A multinational meta-analysis of age differences. *Child Development*, *82*(6), 1715–1737. https://doi.org/10.1111/j.1467-8624.2011.01668.x

Roberts, S. O., & Gelman, S. A. (2016). Can White children grow up to be Black? Children's reasoning about the stability of emotion and race. *Developmental Psychology*, *52*(6), 887–893. https://doi.org/10.1037/dev0000132

Rule, N. O., Ambady, N., Adams, R. B., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*, *98*(1), 1–15. https://doi.org/10.1037/a0017673

Rutland, A., Cameron, L., Milne, A., & McGeorge, P. (2005). Social norms and self-presentation: Children's implicit and explicit intergroup attitudes. *Child Development*, *76*(2), 451–466. https://doi.org/10.1111/j.1467-8624.2005.00856.x

Rutland, A., Killen, M., & Abrams, D. (2010). A new social-cognitive developmental perspective on prejudice. *Perspectives on Psychological Science*, *5*(3), 279–291. https://doi.org/10.1177/1745691610369468

Schug, J., Alt, N. P., & Klauer, K. C. (2015). Gendered race prototypes: Evidence for the non-prototypicality of Asian men and Black women. *Journal of Experimental Social Psychology*, *56*, 121–125. https://doi.org/10.1016/j.jesp.2014.09.012

Sesko, A. K., & Biernat, M. (2010). Prototypes of race and gender: The invisibility of Black women. *Journal of Experimental Social Psychology*, *46*(2), 356–360. https://doi.org/10.1016/j.jesp.2009.10.016

Shutts, K., Banaji, M. R., & Spelke, E. S. (2010). Social categories guide young children's preferences for novel objects. *Developmental Science*, *13*(4), 599–610. https://doi.org/10.1111/j.1467-7687.2009.00913.x

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. New York: Cambridge University Press. https://doi.org/10.2307/2655372

Statistics Canada. (2016). Immigration and ethnocultural diversity highlight tables – Victoria, BC. Retrieved September 15, 2018, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/imm/Table.cfm?Lang=E&T=44&SP=1&geo=59&vismin=5&age=1&sex=1

Suzuki, A. (2018). Persistent reliance on facial appearance among older adults when judging someone's trustworthiness. *Journals of Gerontology – Series B Psychological Sciences and Social Sciences*, *73*(4), 573–583. https://doi.org/10.1093/geronb/gbw034

Suzuki, A., Tsukamoto, S., & Takahashi, Y. (2019). Faces tell everything in a just and biologically determined world: Lay theories behind face reading. *Social Psychological and Personality Science*, *10*(1), 62–72. https://doi.org/10.1177/1948550617734616

Todorov, A. (2017). *Face value : The irresistible infuence of firsts impressions*. Princeton, NJ: Princeton University Press.

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*(4), 724–738. https://doi.org/http://dx.doi.org/10.1037/a0032335

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Oosterhof, N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*(2), 117–122. https://doi.org/10.1109/MSP.2010.940006

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

U.S. Census Bureau. (2016). American Fact Finder – Cambridge, MA. Retrieved September 15, 2018, from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

Weisman, K., Johnson, M. V., & Shutts, K. (2015). Young children's automatic encoding of social categories. *Developmental Science*, *18*(6), 1036–1043. https://doi.org/10.1111/desc.12269

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, *16*(1), 56–63. https://doi.org/10.1111/j.0956-7976.2005.00780.x

Willis, J., & Todorov, A. (2006). First impressions. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325–1331. https://doi.org/10.1177/0956797615590992

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2012). Perception and motivation in face recognition: A critical review of theories of the cross-race effect. *Personality and Social Psychology Review*, *16*(2), 116–142. https://doi.org/10.1177/1088868311418987

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996

Zebrowitz, L. A., Bronstad, P. M., & Montepare, J. M. (2011). An ecological theory of face perception. In In R. B. Adams, N. Ambady, K. Nakayama, & S. Shimojo (Eds.), *The science of social vision* (pp. 3–30). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195333176.003.0002

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497–1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

SUPPLEMENTARY MATERIALS

STUDY 1

For Study 1, the baseline model was a null model containing only a random effect for participants. In subsequent steps we added random intercepts for (1) stimulus set (i.e., the base face) nested within face trait (i.e., trustworthiness, competence, dominance) and (2) order of presentation. The fixed effects were grand-mean centered covariates of gender and race, as well as face pair type, which served as the effect of interest.

Interaction of Face Pair and Trait Type in Selecting the Expected Face. The model with two random effects of subject and stimulus set significantly improved model fit above the previous model with only a random effect of subject, $\chi^2(1) = 207.48$, $p < .001$. A random intercept of order did not improve model fit, $\chi^2(1) = 1.06$, $p = .30$. The two-way fixed effect interaction model between face pair type and face-trait provided significantly better fit than the previous model with only the two main effects, $\chi^2(6) = 38.99$, $p < .001$.

Compared to the dummy-coded baseline of White-White (WW) face pairs in trustworthiness face-trait trials, significant interactions emerged only for Black-Black (BB) face pairs in competence face-trait trials, OR = 2.04, $z = 2.72$, $p = .007$. No other significant interactions emerged, all zs [-0.92, 1.69], all ps > .10. This suggests that face-trait did not consistently and significantly moderate the effect of face-pair type on participants' expected responses. Indeed, inspection of the model predicted percentages reveals the impressive finding that, in nearly all cases, participants are consistently focusing on face-trait cues and selecting the expected face significantly above-chance.

Nevertheless, inspection of the model predicted percentages (Table S1.3) indicates that participants were significantly below-chance in giving the expected face-based response on only one occasion: Black-Negative White-Positive (B-W+) face pairs in competence trials. Indeed, on these trials, participants gave the face-based

TABLE S1.1. Random Effects and Fixed Effects Model Comparisons for Study 1

Random Intercepts	Model *Df*	Model AIC	χ^2	χ^2 *Df*	*p*
Participant (Baseline model)	2	9515.6	--	--	--
+ Stimulus Set in Face Trait	4	8448.5	1071.05	2	< .001
+ Order	5	8449.4	1.11	1	.29
Fixed Effects					
Baseline (Intercept-only)	4	8448.5	--	--	--
+ Covariates (Gender and Race)	6	8451.1	1.49	2	.47
+ Face Pair Type	9	7347.3	1109.71	3	< .001

Note. χ^2 values from likelihood ratio tests testing improvement in model fit as a result of entering new predictors in a stepwise fashion in the mixed-effects logistic regression with selecting the expected face as the dependent variable (coded as 1 for expected). For the supplemental analysis of moderation by face-trait, the baseline model was a null model containing only a random effect for participants. In subsequent steps we added random intercepts for (1) stimulus set (i.e., the base face) and (2) order of presentation. The fixed effects were grand-mean centered covariates of gender and race, as well as face pair type and face-trait, as well as their interaction.

TABLE S1.2. Random Effects and Fixed Effects Model Comparisons, with Moderation by Face-Trait, for Study 1

Random Intercepts	Model *Df*	Model AIC	χ^2	χ^2 *Df*	*p*
Participant (Baseline model)	2	9515.6	--	--	--
+ Stimulus Set	3	9310.1	207.48	1	< .001
+ Order	4	9311.1	1.06	1	.30
Fixed Effects					
Baseline (Intercept-only)	3	9310.1	--		
+ Covariates (Gender and Race)	5	9312.6	1.50	2	.47
+ Face-Trait	7	8506.6	810.00	2	< .001
+ Face Pair Type	8	8348.9	159.66	1	< .001
+ Face Pair Type + Face-Trait	10	7417.9	935.01	2	< .001
+ Face Pair Type x Face-Trait	16	7390.9	38.99	6	< .001

Note. χ^2 values from likelihood ratio tests testing improvement in model fit as a result of entering new predictors in a stepwise fashion in the mixed-effects logistic regression with selecting the expected face as the dependent variable (coded as 1 for expected).

response on only 27% of trials, implying that they were consistently selecting the incompetent-looking *Black* face as "nice" even above the competent-looking *White* face. This result is important as it suggests that in cases where face-trait cues may be deemed irrelevant (e.g., competence appearance is less relevant to niceness judgments), participants may look to other cues (e.g., face-race) to guide their social decisions.

Further, the results from competence trials shows the power of correction for face-race cues: When face-trait cues are perceived as less relevant, participants show an increased tendency to select the Black face as nice (even when that face has an incompetent appearance). Indeed, the largest effects of correction are observed on competence trials, with a gap of 60 percentage points between B+W- and B-W+ face pairs on competence trials but a gap of only 16 percentage points between B+W- and B-W+ face pairs on trustworthiness trials.

STUDY 2

For Study 2, the baseline model was a null model containing only a random effect for participants. In subsequent steps we added random intercepts for (1) stimulus set (i.e., the base face) nested within face trait (i.e., trustworthiness, competence, dominance) and (2) order of presentation.

Additionally, following random effects specification, we compared model fit for the fixed effects of (1) grand-mean centered covariates of gender and race; (2) prompt type (self, other) alone; (3) face pair (WW, BB, B+W-, B-W+) alone; (4) face pair and prompt type as two main effects (face pair + prompt type); and (5) face pair and prompt type as two main effects and their interaction (face pair x prompt type).

TABLE S1.3. Model Predicted Percentages with Moderation by Face-Trait: Study 1

Face-trait	Face Pair							
	White-White		Black-Black		Black-Positive White-Negative		Black-Negative White-Positive	
	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI
Competence	68.70	[56.56%, 78.72%]	76.33	[65.59%, 84.50%]	87.14	[79.82%, 92.06%]	26.86	[17.86%, 38.27%]
Dominance	90.46	[84.59%, 94.25%]	85.41	[77.42%, 90.90%]	92.72	[87.94%, 95.70%]	60.03	[47.17%, 71.64%]
Trustworthiness	96.71	[94.13%, 98.18%]	95.50	[92.21%, 97.44%]	98.17	[96.49%, 99.06%]	82.74	[73.81%, 89.07%]

Note. Results are the predictions from the best-fitting model with random effects of subject and stimuli and fixed effect interaction of face pair and face-trait plus grand-mean centered covariates of gender and race.

Interaction of Face Pair, Prompt Type, and Trait Type in Selecting the Expected Face. The model with two random effects of subject and stimulus set significantly improved model fit above the previous model with only a random effect of subject, $\chi^2(1) = 576.36$, $p < .001$. A random intercept of order did not improve model fit, $\chi^2(1) = 0.003$, $p = .95$. The three-way fixed effect interaction model between face pair type, prompt type (self vs. other) and face-trait provided significantly better fit than the previous models with only the two-way interaction between face pair type and face-trait, $\chi^2(12) = 739.24$, $p < .001$.

Compared to the dummy-coded baseline of White-White (WW) face pairs in the "self" prompt trials and for the trustworthiness face-trait, significant three-way interactions emerged only for Black-Positive White-Negative (B+W-) face pairs in "other" trials for competence face-trait, OR = 1.79, $z = 2.04$, $p = .04$. No other significant interactions emerged, all zs [-0.91, 1.03], all ps > .30. This suggests that face-trait did not consistently and significantly moderate the interaction of face-pair type and prompt on participants' expected responses.

Indeed, inspection of the model predicted percentages (Table S2.3) reveals that, in nearly all cases, participants consistently focus on face-trait cues and select the expected face significantly above-chance. Crucially, inspection of model predictions also shows that the effect of prompt type is consistent across face traits: participants are more likely to select the expected face based on face-trait cues when responding from the perspective on an "other" (as seen in consistent increases in expected responses for B-W+ trials and consistent decreases in expected responses for B+W- trials).

Nevertheless, as in Study 1, inspection of the model predicted percentages indicates that participants were significantly below-chance in giving the expected response on only one occasion: Black-Negative White-Positive (B-W+) face pairs in competence trials for the "self" prompt. This reinforces the previous interpretation that when face-trait cues may be deemed irrelevant, participants preferentially focus on face-race cues in their evaluative judgments. Crucially, the replication in Study 2 also shows that such below-chance responding is exclusively when participants are motivated to correct for the use of face-race cues out of concern for appearing personally prejudiced. Again, this reinforces the powerful mechanism of race-based correction in explaining the observed results.

STUDY 3

For Study 3, the baseline model was a null model containing only a random effect for participant. In subsequent steps we added random intercepts for (1) stimulus set (i.e., the base face) nested within face trait (i.e., trustworthiness, competence, dominance) and (2) judgment type (i.e., mean or nice). The fixed effects were grand-mean centered covariates of gender and race, as well as face pair type, which served as the effect of interest.

Interaction of Face Pair and Trait Type in Selecting the Expected Face. The model with three random effects of subject, stimulus set, and judgment type (mean vs. nice)

significantly improved model fit above the previous model with two random effects of subject and stimuli, $\chi^2(1) = 8.57$, $p = .003$. The two-way fixed effect interaction model between face pair type and face-trait provided significantly better fit than the previous model with only the two main effects, $\chi^2(6) = 219.25$, $p < .001$.

Compared to the dummy-coded baseline of White-White (WW) face pairs in trustworthiness face-trait trials, significant interactions emerged for nearly all face pairs in both dominance and competence face-trait trials. The only non-significant interaction was for Black-Black face pairs in competence face-trait trials, OR = 0.78, $z = -1.53$, $p = .13$. For all other interactions participants were found to be more likely to make expected responses than the dummy-coded baseline: in Black-Black (BB) face pairs for dominance face-trait, OR = 0.13, $z = -9.86$, $p < .001$; in Black-Positive White-Negative (B+W-) face pairs for competence face-trait trials, OR = 1.69, $z = 2.88$, $p = .004$, and for dominance face-trait trials, OR = 0.23, $z = -6.65$, $p < .001$; and in Black-Negative White-Positive (B-W+) face pairs for competence face-trait trials, OR = 0.59, $z = -3.60$, $p < .001$, and for dominance face-trait trials, OR = 0.38, $z = -5.05$, $p < .001$.

This suggests that because participants in WW face pairs in trustworthiness trials were unusually low in their rate of expected responses (given results from other studies), it appears that face-trait significantly moderated the effect of face-pair type on participants' expected responses. However, inspection of the model predicted percentages (Table S3.3) again reveals the consistent finding that, in nearly all cases, participants focus on face-trait cues and select the expected face significantly above-chance.

Indeed, as with Studies 1 and 2, the only occasion of below-chance responses was in Black-Negative White-Positive (B-W+) face pairs in competence trials. The results therefore reinforce that, although perceivers may have a strong baseline tendency to use face-trait cues, there are nevertheless boundary conditions that will sway perceivers to use (and possibly correct for) alternate information such as face-race cues.

ADDITIONAL ANALYSES INCLUDING ALL TRIALS

Analytic Strategy. All trials, including those where the first response was offered after 2500 ms, are retained. As in the main analyses, a generalized linear mixed-effects model (GLMM) was used with the same random effects. Again, the model with all three random effects of subject, stimulus set nested in trait type, and judgment type provided better fit than the previous model with two random effects of subject and stimulus nested in trait, $\chi^2(1) = 16.94$, $p < .001$. The fixed effect of face pair type provided significantly better fit than the previous covariates-only model, $\chi^2(3) = 774.72$, $p < .001$.

Overall Effects. Across all face pair types participants were significantly more likely than chance to select the expected face ($M = 78.83\%$, $SD = 7.40\%$), $t(182) = 52.69$, $p < .001$, $d = 3.89$. No other racial groups differed significantly from White participants ($M_{White} = 78.89\%$, $M_{Black/African\text{-}American} = 76.46\%$, $M_{Asian} = 77.95\%$, and $M_{Other} = 79.17\%$), $R^2 = 0.004$, all $ps > .05$.

TABLE S3.3. Model Predicted Percentages with Moderation by Face-Trait: Study 3

Face-Trait	Face Pair							
	White-White		Black-Black		Black-Positive White-Negative		Black-Negative White-Positive	
	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI
Competence	60.48	[51.47%, 68.84%]	68.76	[60.33%, 76.10%]	86.14	[80.81%, 90.16%]	36.21	[28.21%, 45.07%]
Dominance	95.17	[92.76%, 96.80%]	82.93	[76.88%, 87.66%]	91.56	[87.86%, 94.21%]	82.53	[76.38%, 87.35%]
Trustworthiness	86.11	[80.83%, 90.12%]	91.97	[88.45%, 94.48%]	93.73	[90.81%, 95.76%]	79.48	[72.72%, 84.92%]

Note. Results are the predictions from the best-fitting model with random effects of subject, stimuli, and judgment type (mean vs. nice) and fixed effect interaction of face pair and face-trait plus grand-mean centered covariates of gender and race.

TABLE S3.2. Random Effects Model Comparisons, with Moderation by Face-Trait, for Study 3

Random Intercepts	Model *Df*	Model AIC	χ^2	*Df*	*p*
Participant (Baseline model)	2	13616	--	--	--
+ Stimulus Set	3	13474	143.64	1	< .001
+ Judgment Type	4	13468	8.57	1	.003
Fixed Effects					
Baseline (Intercept-only)	4	13498	--	--	--
+ Covariates (Gender and Race)	6	13470	1.99	2	.37
+ Face-Trait	8	12392	1082.20	2	< .001
+ Face Pair Type	9	12872	0.00	1	> .99
+ Face Pair + Face-Trait	11	11758	1117.74	2	< .001
+ Face Pair x Face-Trait	17	11551	219.25	6	< .001

Note. χ^2 values from likelihood ratio tests testing improvement in model fit as a result of entering new predictors in a stepwise fashion in the mixed-effects logistic regression with selecting the expected face as the dependent variable (coded as 1 for expected).

Effect of Face Pair in Selecting the Expected Face Under Time Constraints (All Trials). As in the main analyses, the effect of face pair was slightly reduced relative to Experiment 1 and first-person (self) trials in Experiment 2, although the influence of correction for race cues again persisted. Compared to White-White face pairs (M_{pred} = 83.47% [66.35%, 92.82%]) participants were equally likely to pick the expected face in Black-Black face pairs (M_{pred} = 83.56% [66.50%, 92.86%]), OR = 1.01, z = 0.12, p = .91. However, compared to WW face pairs, participants were less likely to select the expected face in Black-Negative-White-Positive face pairs (M_{pred} = 68.59% [46.06%, 84.82%]), OR = 0.43, z = -15.66, p < .001, and were more likely to select the expected face in Black-Positive-White-Negative face pairs (M_{pred} = 91.85% [81.45%, 96.66%]), OR = 2.23, z = 12.39, p < .001. As for the main analyses, comparing the effect sizes across Experiments 1 and 3 indicates that the likelihood to pick the expected face in B-W+ face pairs was halved in Experiment 3, suggesting a notable reduction in correction for race bias when participants were under time constraints.

STUDY 4

For Study 4, the baseline model was a null model containing only a random effect for participants. In subsequent steps we added random intercepts for (1) stimulus set (i.e., the base face) nested in face trait (i.e., trustworthiness, competence, dominance), (2) location (Canada or USA), and (3) order of presentation.

Additionally, following random effects specification, we compared model fit for the fixed effects of (1) grand mean-centered covariates of participant gender and race; (2) age (continuous, 5–13 years) alone; (3) face pair (WW, BB, B+W-, B-W+) alone; (4) face pair and age as two main effects (face pair + age); and (5) face pair and age as two main effects and their interaction (face pair x age).

TABLE S4.1. Random Effects and Fixed Effects Model Comparisons for Study 4

Random Intercepts	Model *Df*	Model AIC	χ^2	*Df*	*p*
Participant (Baseline model)	2	14671	--	--	--
+ Stimulus Set in Face Trait	4	13603	1071.80	2	< .001
+ Location	5	13605	0.00	1	> .99
+ Order	6	13607	0.00	1	> .99
Fixed Effects					
Baseline (Intercept-only)	4	13603	--	--	--
+ Covariates (Gender and Race)	6	13601	6.45	2	.04
+ Age	7	13595	7.31	1	.007
+ Face Pair	9	13248	351.64	2	< .001
+ Age + Face Pair	10	13242	7.30	1	.007
+ Age x Face Pair	13	13174	74.52	3	< .001

Note. As before, χ^2 values from likelihood ratio tests testing improvement in model fit as a result of entering new predictors in a stepwise fashion in the mixed-effects logistic regression with selecting the expected face as the dependent variable (coded as 1 for expected).

Interaction of Face Pair, Age and Trait Type in Selecting the Expected Face. The model with two random effects of subject and stimulus set significantly improved model fit above the previous model with only a random effect of subject, $\chi^2(1) = 208.94$, $p < .001$. No other random intercepts improved model fit, all χ^2s < 0.01, $p > .99$. The three-way fixed effect interaction model between face pair type, age, and face-trait provided significantly better fit than the previous models with only the two-way interaction between face pair type and face-trait, $\chi^2(12) = 97.45$, $p < .001$.

Despite this significantly better fit, only one significant three-way interaction emerged when comparing to the dummy-coded baseline of White-White (WW) face pairs in trustworthiness face-trait, although the effect was small. Specifically, there was a significant interaction of age, face pair and face-trait for Black-Positive White-Negative (B+W-) face pairs in dominance face-trait trials, OR = 1.18, $z = 2.15$, $p = .03$. No other three-way interactions were significant, all zs [0.19, 1.55], all ps > .12. This suggests that, as with the previous studies reported above, face-trait did not moderate the interaction of face-pair type and age on participants' expected responses.

Indeed, the model predicted percentages (Table S4.3) again reveal that participants consistently focus on face-trait cues and select the expected face above-chance in nearly every case. Additionally, inspection of model predictions shows that the effect of age is consistent across face traits: the youngest participants show little variation across face pair types, implying little motivation to correct for face-race cues, while the oldest participants show much larger differences across B-W+ and B+W- face pairs, suggesting developing motivations to correct for face-race cues.

Finally, it is notable that the lowest rate of expected responses again emerged in Black-Negative White-Positive (B-W+) face pairs in competence trials, although only for the oldest children. Thus, across all four studies, exploring moderation by

TABLE S4.2. Random Effects and Fixed Effects Model Comparisons, with Moderation by Face Trait, for Study 4

Random Intercepts	Model *Df*	Model AIC	χ^2	*Df*	*p*
Participant (Baseline model)	2	14671	--	--	--
+ Stimulus Set	3	14464	208.94	1	< .001
+ Location	4	14466	0.00	1	> .99
+ Order	5	14468	0.00	1	> .99
Fixed Effects					
Baseline (Intercept-only)	3	14464	--	--	--
+ Covariates (Gender and Race)	5	14462	6.20	2	.045
+ Age	6	14456	7.17	1	.007
+ Face-Trait	7	13632	826.05	1	< .001
+ Face Pair	8	14136	0.00	1	> .99
+ Age + Face-Trait	8	13627	508.16	0	< .001
+ Age + Face Pair	9	14130	0.00	1	> .99
+ Face-trait + Face Pair	10	13282	850.17	1	< .001
+ Age x Face-Trait	10	13627	0.00	0	> .99
+ Age + Face Pair + Face-Trait	11	13277	352.05	1	< .001
+ Age x Face Pair	12	14066	0.00	1	> .99
+ Face-Trait x Face Pair	16	13260	814.73	4	< .001
+ Age x Face Pair x Face-Trait	28	13186	97.45	12	< .001

Note. As before, χ^2 values from likelihood ratio tests testing improvement in model fit as a result of entering new predictors in a stepwise fashion in the mixed-effects logistic regression with selecting the expected face as the dependent variable (coded as 1 for expected).

face-trait reinforces the conclusion that when face-trait cues are less relevant or diagnostic (e.g., competence appearance cues for niceness judgments), participants may preferentially focus on face-race cues in their evaluative judgments. Crucially, it appears that participants will only use (and correct for) such information if they are old enough to both understand the implications of race as a social category and have sufficient motivation to correct for possible race-based bias.

TABLE S4.3. Model Predicted Percentages with Moderation by Face-Trait: Study 4

Face-Trait x Age	Face Pair							
	White-White		Black-Black		Black-Positive White-Negative		Black-Negative White-Positive	
	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI	Mean (%)	95% CI
Competence (youngest kids)	64.48	[53.40%, 74.21%]	60.76	[49.49%, 71.00%]	55.28	[43.69%, 66.33%]	63.56	[52.53%, 73.32%]
Dominance (youngest kids)	83.59	[75.61%, 89.33%]	73.96	[63.94%, 81.99%]	82.32	[73.83%, 88.49%]	67.26	[56.51%, 76.47%]
Trustworthiness (youngest kids)	90.09	[83.58%, 94.20%]	89.18	[82.60%, 97.54%]	83.51	[74.86%, 98.50%]	87.66	[81.07%, 87.93%]
Competence (oldest kids)	75.90	[66.10%, 83.57%]	74.28	[64.20%, 82.29%]	91.41	[86.39%, 94.69%]	46.55	[35.34%, 58.12%]
Dominance (oldest kids)	86.47	[79.16%, 91.49%]	81.00	[72.26%, 87.47%]	91.25	[85.81%, 94.74%]	62.68	[51.22%, 72.88%]
Trustworthiness (oldest kids)	97.52	[95.08%, 98.76%]	95.49	[91.88%, 97.54%]	97.15	[94.66%, 98.50%]	81.40	[72.43%, 87.93%]

Note. Results are the predictions from the best-fitting model with random effects of subject and stimuli and fixed effect three-way interaction of face pair, age (continuous), and face-trait plus grand-mean centered covariates of gender and race.