Supplementary Analyses and Detail: Table of Contents

Experiments 1a and 1b

Experiment 2b

Experiments 1a and 1b

Additional information on Experiments 1a and 1b is included below. Please also refer to the .qsf files on Qualtrics for any questions about how the surveys themselves were built.

Conversation Topic Pretesting

Experiments 1a and 1b both required participants to watch a conversation unfold that was specifically normed to be neutral with respect to the target identities in each experiment. In Experiment 1a, the conversation topics were normed to ensure neutrality with respect to targets' age and gender groups; in Experiment 1b, the conversation topics were normed to ensure neutrality with respect to targets' race and gender groups.

Experiment 1a Pretesting

Sample and method. A sample MTurk workers ($N = 56$) pre-tested a series of conversation stances (e.g., zoos are unethical, Americans drink too much caffeine, etc.)—all taken from a high-school debate team website—on whether men *vs.* women would agree with them more (from *1 = men would agree more* to *7 = women would agree more*), and on whether young *vs.* older adults would agree with them more (from *1 = young adults would agree more* to *7 = older adults would agree more*).

Pretesting results. Of the stances that were pretested, two were chosen based on these ratings. According to these ratings, the stance "Introverts are more cerebral than extraverts" was no more associated with men than with women (according to a one-sample *t*-test: $t(55) = 1.32$, $p = .19$), and it was associated no more with older adults than with young adults (according to a one-sample *t*-test: $t(55) = -0.29$, $p = .77$). In addition, the stance "People who mirror whomever they're talking to are inauthentic" was no more associated with men than with women (one-sample *t*-test: $t(54) = 0.43$, $p = .67$), and it was no more associated with older adults than with young adults (one-sample *t*-test: $t(55) = -0.49$, $p = .63$).

Experiment 1b Pretesting

Sample and method. A sample MTurk workers ($N = 54$) pre-tested a series of conversation stances (e.g., everyone should be vegetarian, obesity is a disease, etc.)—all taken from a high-school debate team website—on whether men *vs.* women would agree with them more (from *1 = men would agree more* to *7 = women would agree more*), and on whether White *vs.* Black Americans would agree with them more (from *1 = White Americans would agree more* to *7 = Black Americans would agree more*).

Pretesting results. Of the stances that were pretested, two were chosen based on these ratings. According to these ratings, the stance "Committing suicide should be made illegal" was no more associated with men than with women (according to a one-sample *t*-test: $t(53) = 0.89$, $p = .38$), and it was associated no more with White Americans than with Black Americans (according to a one-sample *t*-test: $t(52) = 0.57$, $p = .57$). In addition, the stance "Celebrities earn too much money" was no more associated with men than with women (one-sample *t*-test: *t*(52) = 0.81, *p* = .42), and it was no more associated with White Americans than with Black Americans (one-sample *t*-test: *t*(52) = -0.34, *p* = .74).

Checklist Trait Ratings

In Experiments 1a and 1b, one of the supplementary measures required participants to nominate stereotypic attributes, from a checklist of 99 attributes, about a randomly assigned target person (an older women from the conversation they had viewed, in Experiment 1a; a Black woman from the conversation they had viewed, in Experiment 1b). The 99 traits that we included in our checklist were adapted directly from Petsko & Bodenhausen, 2019 (JESP), as were the ratings of stereotypic “Blackness” and “femininity.” However, for Experiment 1a, we needed to have the checklist attributes rated on stereotypic “oldness” as well. To gather these ratings, we followed the exact same procedure as outlined in Petsko & Bodenhausen, 2019. Details on oldness ratings are below.

Oldness Ratings

Sample and method. A sample of MTurk workers (*N* = 81) rated all 99 checklist traits on the degree to which they seem stereotypic of old people. Of these participants, I excluded *n* = 4 (4.94%) for not responding “yes” to the question, “Did you take this survey seriously?” Participants rated all 99 traits, in a randomized order on how stereotypic they seemed, from *1 = not at all* to *7 = very much.*

Oldness ratings results. Average “oldness” scores were created for each attribute, which were then imputed into participants’ trait selections in Experiment 1a. For illustrative purposes the “oldest” 10 words in the checklist were, in order: *loyal to family ties, tradition-loving, stubborn, conservative, conventional, very religious, faithful, straightforward, honest,* and *practical*. The 10 “least old” words in the checklist were, from “least” to “more old”: *athletic, criminal, violent, sexually perverse, mercenary, sensual, progressive, radical, treacherous, gluttonous,* and *aggressive*.

How Internal Replication Factors Influence Results

Our manuscript reports that in both Experiment 1a and 1b, participants saw a) one of several possible sets of stimulus faces and b) one of two different conversation topics during the who-said-what task. Both of these factors were meant to serve as internal replication factors only, and were meant to ensure generalizability of the findings. Because the overarching interpretation of our findings do not depend on these factors, we simply collapse across them in our analyses. Here, we describe them in greater detail.

In general, which stimulus set of faces participants were assigned accounted for virtually no error in participants’ responses on the who-said what task (in Experiment 1a, stimulus face set accounted for less than .001% of the variance in outcomes; in Experiment 1b, stimulus face set also accounted for less than .001% of the variance in outcomes). Thus, we dropped the random effect of intercept of stimulus face set from these models in our paper. Including these random effects leave our results virtually unchanged.

That said, which conversation topic participants were assigned did occasionally influence the magnitude of the results in both Experiments 1a and 1b. Notably, these influences never changed the overarching interpretation of results.

Experiment 1a: Influence of Conversation Topic

Who-said-what: Age categorization. In the manuscript, age categorization is assessed by analyzing participants' errors in a 2(error type: within-age-group, between-age-group) × 3 (condition: age fit, control gender fit) mixed ANOVA with repeated measures on the first factor. Here, these same errors are analyzed as a 2(error type: within-age-group, between-age-group) × 3 (condition: age fit, control, gender fit) × 2 (conversation topic: introversion, phoniness) mixed ANOVA with repeated measures on the first factor.

This analysis reveals the exact same pattern of results reported in the manuscript: a main effect of error type suggesting that participants are engaging in age categorization, on average ($M_{\text{diff}} = 4.67$, 95% CI[4.07, 5.26], $\beta = 0.93$, $F(1, 295) = 235.04$, $p < .001$); and an interaction between error type and fit condition ($F(1, 295) = 235.93$, $p < .001$, $\omega_p^2 = 0.44$). The interaction was almost identical in magnitude as that reported in the manuscript and was not moderated by which conversation topic participants were assigned to see ($F(1, 295) = 0.03$, $p = .86$, $\omega_p^2 < 0.01$). The only influence of conversation topic on participants' responses was a main effect: participants made slightly more total errors in the who-said-what task in the introversion conversation condition than in the phoniness conversation condition: $M_{\text{diff}} = 0.64$, 95% CI[0.02, 1.25], $\beta = 0.13$, $F(1, 295) = 4.15$, $p = .042$.

Who-said-what: Gender categorization. In the manuscript, gender categorization is assessed by analyzing participants' errors in a 2(error type: within-gender-group, between-gender-group) × 3 (condition: age fit, control gender fit) mixed ANOVA with repeated measures on the first factor. Here, these same errors are analyzed as a 2(error type: within-gender-group, between-gender-group) × 3 (condition: age fit, control, gender fit) × 2 (conversation topic: introversion, phoniness) mixed ANOVA with repeated measures on the first factor.

This analysis reveals the exact same pattern of results reported in the manuscript: a main effect of error type suggesting that participants are engaging in gender categorization, on average ($M_{\text{diff}} = 5.09$, 95% CI[4.51, 5.67], $\beta = 0.99$, $F(1, 295) = 294.02$, $p < .001$); and an interaction between error type and fit condition ($F(1, 295) = 215.41$, $p < .001$, $\omega_p^2 = 0.42$). The nature of this interaction was that there was substantially greater gender categorization in the gender-fit condition relative to the other two conditions. This interaction, however, *was* different in magnitude depending on which conversation topic participants were assigned ($F(1, 295) = 21.96$, $p < .001$, $\omega_p^2 = 0.07$). Deconstructing this interaction reveals that the influence of the gender-fit condition (*vs.* the other two conditions) was weaker when the conversation topic was about introversion [$F(1, 295) = 48.61$, $p < .001$, $\omega_p^2 = 0.14$] than when it was about the phoniness of self-monitors [$F(1, 295) = 192.60$, $p < .001$, $\omega_p^2 = 0.39$].

Experiment 1b: Influence of Conversation Topic

Who-said-what: Race categorization. In the manuscript, race categorization is assessed by analyzing participants' errors in a 2(error type: within-race-group, between-race-group) × 3 (condition: race fit, control gender fit) mixed ANOVA with repeated measures on the first factor. Here, these same errors are analyzed as a 2(error type: within-race-group, between-race-group) × 3 (condition: race fit, control, gender fit) × 2 (conversation topic: celebrities, suicide) mixed ANOVA with repeated measures on the first factor.

This analysis yields the two same results reported in the manuscript. Namely, across conditions there is a pronounced tendency for participants to engage in race categorization [M_{diff} = 4.00, 95% CI[3.39, 4.61], β = 0.78, $F(1, 568) = 168.01$, $p < .001$], and the degree of race categorization depends on whether or not participants were in the race-fit condition [$F(1, 568) = 348.94$, $p < .001$, $\omega_p{}^2 = 0.38$]. However, the magnitude of the amount by being in the race-fit condition amplified race categorization was contingent on which conversation topic participants were assigned [$F(1, 568) = 8.29$, $p = .004$, $\omega_p{}^2 = 0.01$]. The nature of this interaction was that the influence of race-fit (vs. other conditions) on participants' race-categorization was slightly weaker in the celebrity topic condition [$F(1, 568) = 121.33$, $p < .001$, $\omega_p{}^2 = 0.17$] than in the suicide topic condition [$F(1, 568) = 239.32$, $p < .001$, $\omega_p{}^2 = 0.30$].

Who-said-what: Gender categorization. In the manuscript, gender categorization is assessed by analyzing participants' errors in a 2(error type: within-gender-group, between-gender-group) × 3 (condition: race fit, control gender fit) mixed ANOVA with repeated measures on the first factor. Here, these same errors are analyzed as a 2(error type: within-gender-group, between-gender-group) × 3 (condition: race fit, control, gender fit) × 2 (conversation topic: celebrities, suicide) mixed ANOVA with repeated measures on the first factor.

This analysis yields the two same results reported in the manuscript. Across conditions there is a pronounced tendency for participants to engage in gender categorization [M_{diff} = 6.13, 95% CI[5.53, 6.74], β = 1.12, $F(1, 284) = 399.24$, $p < .001$], and the degree of gender categorization depends on whether or not participants were in the gender-fit condition [$F(1, 284) = 233.33$, $p < .001$, $\omega_p{}^2 = 0.45$]. However, the magnitude of the amount by being in the gender-fit condition amplified gender categorization was contingent on which conversation topic participants were assigned [$F(1, 284) = 22.98$, $p < .001$, $\omega_p{}^2 = 0.07$]. The nature of this interaction was that the influence of gender-fit (vs. other conditions) on participants' gender-categorization was weaker in the celebrity topic condition [$F(1, 284) = 55.16$, $p < .001$, $\omega_p{}^2 = 0.16$] than in the suicide topic condition [$F(1, 284) = 200.57$, $p < .001$, $\omega_p{}^2 = 0.41$].

Detail on Supplementary DVs

Both Experiments 1a and 1b required participants to a) nominate checklist traits about their targets that were normed on stereotypic "oldness," "Blackness," and "femininity," and to b) rate the faces of individual target old women (Experiment 1a) and Black women (Experiment 1b) from the conversation they had viewed on these same dimensions. These analyses revealed nothing but null effects. Full detail on these (null) analyses are reported below.

Experiment 1a—Supplementary DVs

Checklist attributes: Stereotypic oldness. Participants were expected to nominate traits for older women that were rated as "older" in the age-fit condition, when the lens of *age* was activated, than when in the control or gender-fit conditions. To examine this, the "oldness" of participants' trait nominations was subjected to a one-way ANOVA. Contradicting hypotheses, this analysis yielded null results. Participants' trait attributions were no "older" in the age-fit condition ($M = 4.05$, $SE = 0.04$) than in the control or gender-fit conditions ($M = 4.02$, $SE = 0.03$), $M_{\text{diff}} = 0.03$, 95% CI[–0.07, 0.14], $\beta = 0.07$, $F(1, 292) = 0.39$, $p = .53$. Moreover, these latter conditions did not differ from each other: $M_{\text{diff}} = –0.10$, 95% CI[–0.22, 0.02], $\beta = –0.24$, $F(1, 292) = 2.76$, $p = .10$ (see Figure S1). Thus, participants did not nominate stereotypes for older women that were any "older" in the age-fit condition than in the other conditions.

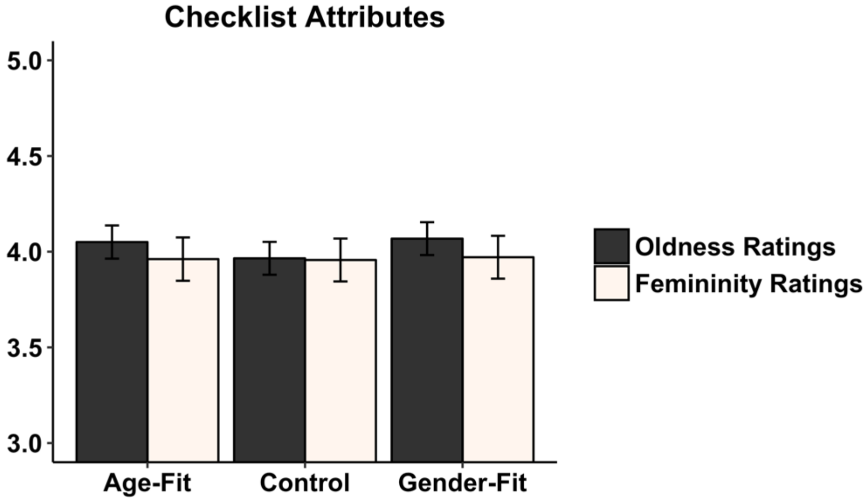

Figure S1. How "old" and "feminine" trait nominations for older women were rated to be (Exp. 1a), broken down by whether participants were in conditions that emphasized the fit of age (left), gender (right), or neither age nor gender (middle) categories. Means are encompassed by 95% confidence intervals.

Checklist attributes: Stereotypic femininity. Participants were expected to nominate "more feminine" attributes for older women in the gender-fit condition than in the other two conditions. Subjecting the stereotypic femininity of participants' nominations to a one-way ANOVA, however, revealed that this was not the case. Participants' trait nominations for older women were no more or less "feminine" in the gender-fit condition ($M = 3.97$, $SE = 0.05$) than in the other two conditions ($M = 3.96$, $SE = 0.04$), $M_{\text{diff}} = 0.01$, 95% CI[–0.13, 0.15], $\beta = 0.02$, $F(1, 292) = 0.03$, $p = .86$. Furthermore, these other two conditions did not differ from each other, $M_{\text{diff}} > –0.01$, 95% CI[–0.16, 0.15], $\beta = –0.01$, $F(1, 292) < 0.01$, $p = .95$ (see Figure S1).

Face ratings: Typical of older adults. Participants were expected to rate the faces of older women as “older looking” when they were in the age-fit condition relative to the other two conditions. A one-way ANOVA indicated, however, that this was not the case: older women were rated as looking no “older” in the age-fit condition (*M* = 6.79, *SE* = 0.17) than in the control or gender-fit conditions (*M* = 7.07, *SE* = 0.12), M_{diff} = –0.28, 95% CI[–0.69, 0.14], β = –0.16, *F*(1, 292) = 1.73, *p* = .19. In addition, ratings in these latter conditions did not differ from each other, M_{diff} = –0.26, 95% CI[–0.73, 0.21], β = –0.15, *F*(1, 292) = 1.18, *p* = .28 (see Figure S2). Thus, in contrast to ICT, there was no evidence that older women were rated as looking “older” in the age-fit condition relative to the other conditions.

Face ratings: Typical of women. Finally, participants were expected to regard older women as looking more typical of women in the gender-fit condition than in the other two conditions. Yet a one-way ANOVA on typicality ratings revealed that this was not the case. Participants did not rate older women as looking more typical of women in the gender-fit condition (*M* = 5.61, *SE* = 0.20) than in the other two conditions (*M* = 5.48, *SE* = 0.14), M_{diff} = 0.13, 95% CI[–0.35, 0.60], β = 0.06, *F*(1, 292) = 0.27, *p* = .61. Moreover, ratings in these other two conditions did not differ from each other, M_{diff} = –0.17, 95% CI[–0.73, 0.38], β = –0.09, *F*(1, 292) = 0.38, *p* = .54 (see Figure S2).

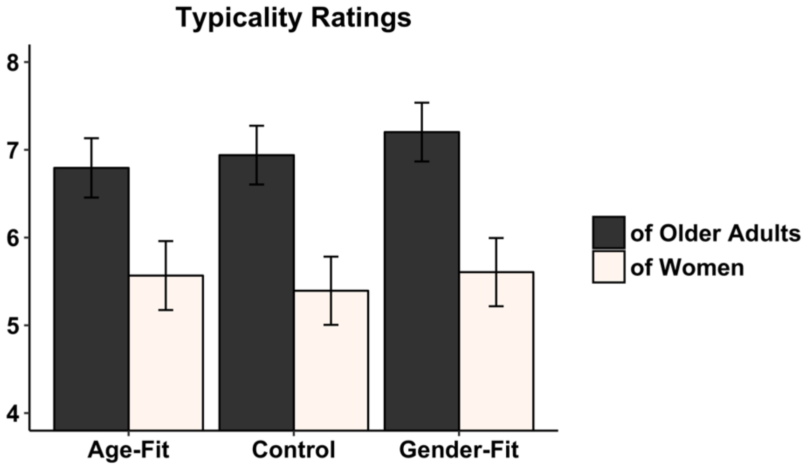

Figure S2. Typicality ratings of old women’s faces (Exp. 1a), broken down by whether participants were in conditions that emphasized the fit of age (left), gender (right), or neither age nor gender (middle) categories. Means are encompassed by 95% confidence intervals.

Experiment 1b—Supplementary DVs

Checklist attributes: Stereotypic blackness. Subjecting the average “Blackness” ratings of participants’ trait nominations to a one-way ANOVA yielded null results. Contrary to hypotheses, participants characterized Black women as no more stereotypically Black in the race-fit condition (*M* = 4.53, *SE* = 0.07) than in the control or gender-fit conditions (*M* = 4.52, *SE* = 0.04), M_{diff} = 0.01, 95% CI[−0.16, 0.18], β = 0.02, *F*(1, 280) = 0.02, *p* = .89. In addition, Black women were characterized by traits that were equally “Black” across these latter conditions, $M_{\text{diff}} = -0.05$, 95% CI[−0.24, 0.18], $\beta = -0.07$, $F(1, 280) = 0.27$, $p = .61$ (see Figure S3).

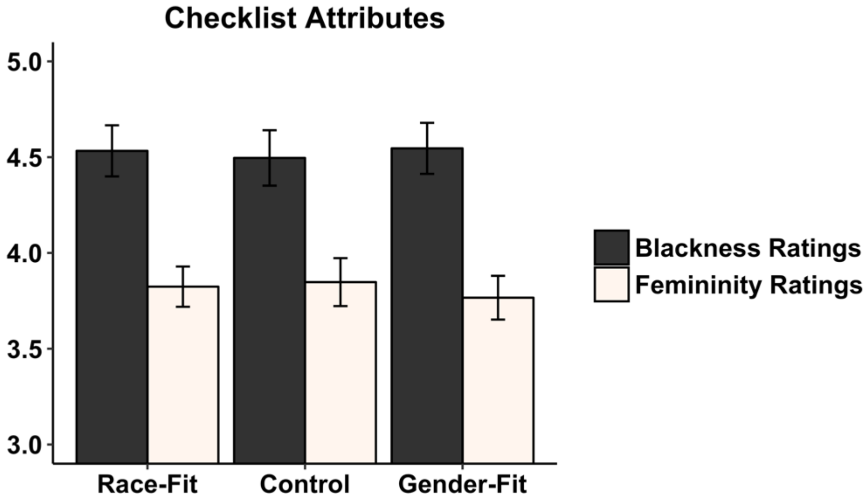

Figure S3. How “Black” and “feminine” trait nominations for Black women were rated to be (Exp. 1b), as a function of which “fit” condition participants had been in: race (left), gender (right), or neither race nor gender (middle). Means are encompassed by 95% confidence intervals.

Checklist attributes: Stereotypic femininity. Subjecting the average “femininity” ratings of participants’ trait nominations to a one-way ANOVA also yielded null results. Contrary to hypotheses, participants characterized Black women as no more stereotypically feminine in the gender-fit condition ($M = 3.77$, $SE = 0.06$) than in the control or race-fit conditions ($M = 3.83$, $SE = 0.04$), $M_{\text{diff}} = -0.07$, 95% CI[−0.21, 0.07], $\beta = -0.12$, $F(1, 280) = 0.97$, $p = .33$. In addition, Black women were characterized by traits that were equally “feminine” regardless of whether they were from the control condition or the race-fit condition, $M_{\text{diff}} = 0.02$, 95% CI[−0.14, 0.19], $\beta = 0.04$, $F(1, 280) = 0.08$, $p = .77$ (see Figure S3).

Face ratings: Typical of Black Americans. A one-way ANOVA indicated that this was not the case: Black women were rated as looking no “Blacker” in the race-fit condition ($M = 6.43$, $SE = 0.19$) than in the control or gender-fit conditions ($M = 6.15$, $SE = 0.13$), $M_{\text{diff}} = 0.29$, 95% CI[–0.15, 0.74], $\beta = 0.17$, $F(1, 281) = 1.69$, $p = .19$. In addition, face ratings of “Blackness” did not vary across the control *vs*. age-fit conditions, $M_{\text{diff}} = 0.03$, 95% CI[–0.48, 0.53], $\beta = 0.02$, $F(1, 292) = 0.01$, $p = .91$ (see Figure S4).

Face ratings: Typical of women. Finally, a one-way ANOVA of how “typical of women” participants rated the Black women also yielded null results. Participants did not rate Black women as looking more typical of women in the gender-fit condition ($M = 5.68$, $SE = 0.19$) than in the other conditions ($M = 5.88$, $SE = 0.14$), [$M_{\text{diff}} = –0.19$, 95% CI[–0.66, 0.28], $\beta = –0.10$, $F(1, 281) = 0.66$, $p = .42$]. Likewise, face ratings of Black women did not vary across the control *vs*. race-fit conditions [$M_{\text{diff}} < 0.01$, 95% CI[–0.55, 0.56], $\beta < 0.01$, $F(1, 281) < 0.01$, $p = .99$] (see Figure S4).

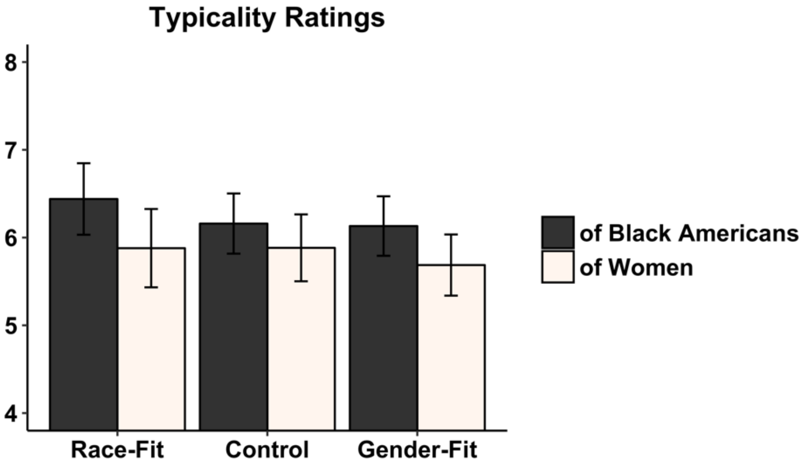

Figure S4. Typicality ratings of Black women's faces (Exp. 1b), broken down by whether participants were in conditions that emphasized the fit of race (left), gender (right), or neither race nor gender (middle) categories. Means are encompassed by 95% confidence intervals.

Exp. 1b Results as a Function of Sample Type (Undergraduate, MTurk)

In footnote 3 of the manuscript, we mention that the effects of interest in Experiment 1b held regardless of whether the participants were sourced from an undergraduate research pool vs. MTurk. However, we also mentioned that there was suggestive evidence that the effects of interest may be more pronounced among undergraduates than among MTurk workers. Below report on the extent to which race categorization and gender categorization, respectively, varied as a function of participant type.

Who-said-what: Race Categorization. In the manuscript, race categorization is assessed by analyzing participants' errors in a mixed linear model that was equivalent to a 2 (error type: within-race-group, between-race-group) × 3 (condition: race fit, control, gender fit) mixed ANOVA with repeated measures on the first factor. Here, these same errors are analyzed as a 2 (error type: within-race-group, between-race-group) × 3 (condition: race fit, control, gender fit) × 2 (sample type: Undergraduate, MTurk) mixed linear model with repeated measures on the first factor.

This analysis yields the two same results reported in the manuscript. Namely, across conditions there is a pronounced tendency for participants to engage in race categorization [$M_{\text{diff}} = 4.01$, 95% CI[3.41, 4.61], $\beta = 0.78$, $F(1, 568) = 170.43$, $p < .001$], and the degree of race categorization depends on whether or not participants were in the race-fit condition [$F(1, 568) = 354.08$, $p < .001$, $R^2 = 0.38$]. However, the magnitude of the amount by being in the race-fit condition amplified race categorization was contingent on whether participants came from the undergraduate sample vs. the MTurk sample [$F(1, 568) = 9.22$, $p = .003$, $R^2 = 0.02$]. The nature of this interaction was that the influence of race-fit (vs. other conditions) on participants' race-categorization was more extreme when the participants were undergraduates [$F(1, 568) = 261.95$, $p < .001$, $R^2 =$

0.32] than when participants were MTurk workers [$F(1, 568) = 104.49, p < .001, R^2 = 0.16$]. Notably, however, the pattern of reported findings was exactly the same independent of the sample. That is, both the MTurk and the undergraduate sample exhibited high levels of race categorization in the race-fit condition (MTurk: $\beta = 1.97, p < .001$; Undergraduate: $\beta = 2.74, p < .001$), and virtually non-existent levels of race categorization in the other two conditions (MTurk: $\beta = -0.01, p = .89$; Undergraduate: $\beta = -0.03, p = .76$). Thus, in summary the effects of interest were present among both groups of participants, but were larger when the participants were undergraduates than when they were MTurk workers.

Who-said-what: Gender Categorization. In the manuscript, gender categorization is assessed by analyzing participants' errors in a mixed linear model that was equivalent to a 2 (error type: within-gender-group, between-gender-group) × 3 (condition: race fit, control, gender fit) mixed ANOVA with repeated measures on the first factor. Here, these same errors are analyzed as a 2 (error type: within-gender-group, between-gender-group) × 3 (condition: race fit, control, gender fit) × 2 (sample type: Undergraduate, MTurk) mixed model with repeated measures on the first factor.

This analysis yields the two same results reported in the manuscript. Namely, across conditions there is a pronounced tendency for participants to engage in gender categorization [$M_{\text{diff}} = 6.13$, 95% CI[5.52, 6.75], $\beta = 1.12, F(1, 568) = 380.56, p < .001$], and the degree of gender categorization depends on whether or not participants were in the gender-fit condition [$F(1, 568) = 221.62, p < .001, R^2 = 0.28$]. However, there was marginal evidence that the magnitude by which by being in the gender-fit condition amplified gender categorization was contingent on whether participants came from the undergraduate sample vs. the MTurk sample [$F(1, 568) = 3.58, p = .059, R^2 = 0.01$]. The nature of this interaction was that the influence of gender-fit (vs. other conditions) on participants' gender-categorization was more extreme when the participants were undergraduates [$F(1, 568) = 151.41, p < .001, R^2 = 0.21$] than when participants were MTurk workers [$F(1, 568) = 73.74, p < .001, R^2 = 0.11$]. Notably, however, the pattern of reported findings was exactly the same independent of the sample. That is, both the MTurk and the undergraduate sample exhibited high levels of gender categorization in the race-fit condition (MTurk: $\beta = 2.01, p < .001$; Undergraduate: $\beta = 2.58, p < .001$), and substantially lower levels of gender categorization in the other two conditions (MTurk: $\beta = 0.46, p < .001$; Undergraduate: $\beta = -0.57, p < .001$). Thus, the effects of interest were present among both groups of participants, but were larger when the participants were undergraduates than when they were MTurk workers.

Experiment 2b

Additional information on Experiment 2b is included below. Please also refer to the .qsf file on OSF for any questions about how the survey itself was built.

Results Using Pre-Registered Exclusion Criterion

In Experiment 2b, we pre-registered our intention to exclude participants who were unusually fast or slow in their response times on the IAT. The criterion we pre-registered was 3 median absolute deviations (MADs) from the median of participants' average response latencies on a given IAT. Thus, if a participant was more than three MADs too fast or too slow—relative to the median of participants' mean response latencies—we planned on eliminating them.

However, when we used this exclusion criterion, we had reason to believe it was not strict enough, and that there were still outliers in our data set that were adding noise to our effect estimates. For example, after excluding participants beyond 3 MADs, we found that skew was substantially higher in Experiment 2b (skew = 41.85) than it had been in Experiment 2a (skew = 8.66). In light of this issue, we decided to use two MADs as the cutoff in Experiment 2b rather than three MADs. Doing so not only reduced the skew in response latencies by a whopping 86.59% (from skew = 41.85 to skew = 5.61), but also reduced the standard error estimates around IAT effects (e.g., of a race-weapons bias) by nearly half (48.53%), and it reduced the standard error estimates around effect-by-IAT-type interaction estimates by nearly half (48.53%) as well. In short, then, there were substantial, data-driven reasons to use two rather than three MADs as our cutoff point for being too fast or too slow in response latency in Experiment 2b.

As noted in our manuscript, significance levels of our statistical tests from Experiment 2b do change when using two rather than three MADs as an exclusion criterion. However, the general pattern of results (reported below) remains the same.

Results when using 3-MAD cutoff. To analyze the data in Experiment 2b (with a three- rather than two-MAD cutoff), response latencies (in milliseconds) for facial stimuli were regressed, in a multilevel model, onto within-person contrast codes that represented the full 2 (IAT type: race-IAT, age-IAT) × 2 (race pairing: Black-weapon + White-harmless; White-weapon + Black-harmless) × 2 (age pairing: adult-weapon + child-harmless; child-weapon + adult-harmless) within-person factorial design of this experiment. This model included estimates of three random effects: a random effect of IAT block intercept, which adjusted for any variation in response latencies that was attributable to some blocks coming earlier in the experiment than others; a random effect of participant intercept, which adjusted for the fact that the full factorial design of this experiment was nested within person; and a random effect of stimulus intercept, which adjusted for the fact that observations were also nested within particular stimulus faces.

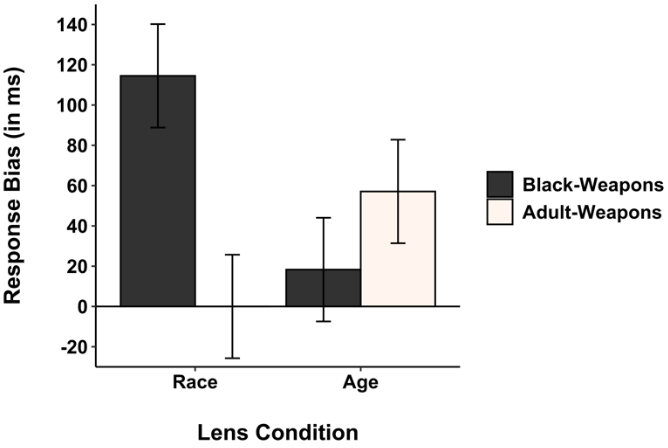

Figure S5. Average Black-weapons (dark gray) and adult-weapons (light pink) implicit associations (Exp. 2b), broken down by whether participants were completing a race-lens IAT (left) or an age-lens IAT (right). Higher scores indicate stronger implicit associations (in milliseconds). Effect estimates are encompassed by 95% confidence intervals. Note that these results differ slightly from those reported in Figure 4 of the manuscript because these results use a more lenient exclusion criterion.

This 2 (IAT type) × 2 (race pairing) × 2 (age pairing) analysis described above revealed a main effect of race pairing. The nature of this main effect was that participants were indeed faster to associate Black individuals with weapons (and White individuals with harmless objects: M = 773.89, SE = 18.62) than the reverse of these pairings (M = 840.29, SE = 18.61), M_{diff} = –66.40ms, 95% CI[–84.58, –48.22], β = –0.11, F(1, 16890) = 51.24, p < .001, R^2 < .01. Thus, participants did indeed exhibit a tendency to implicitly associate Black individuals (both adult men and young boys) with weapons more quickly than they associated White individuals (both adult men and young boys) with weapons. Moreover, and in support of the second hypothesis, this tendency was moderated by condition: β = –0.16, F(1, 16889) = 26.89, p < .001, R^2 < .01. The nature of this interaction was that the tendency to implicitly associate Black individuals (more than White individuals) with weapons was present when faces were being categorized by their racial groups [M_{diff} = –114.50, 95% CI[–140.18, –88.81], β = –0.19, F(1, 16878) = 76.34, p < .001, R^2 < .01], but not when faces were being categorized by their age groups [M_{diff} = –18.30, 95% CI[–44.03, 7.43], β = –0.03, F(1, 16900) = 1.94, p = .16, R^2 < .01]. Thus, results using the 3-MAD cutoff generally replicate results using the 2-MAD cutoff reported in our manuscript (see Figure S5).

In addition, this analysis revealed a main effect of age pairing, such that participants tended to associate adults with weapons (and children with harmless objects: M = 792.82, SE = 18.62) more quickly than the reverse of these pairings (M = 821.36, SE = 18.61), M_{diff} = –28.54ms, 95% CI[–46.71, –10.37], β = –0.05, F(1, 16887) = 9.47, p = .002, R^2 < .01. In addition, and consistent with the results reported in our manuscript, the magnitude of this effect was significantly moderated by IAT type: β = 0.09, F(1, 16890) = 9.48, p = .002, R^2 < .01. The nature of this interaction was that the tendency to associate adults with weapons (and children with harmless objects) emerged strongly when participants were categorizing targets by age: M_{diff} = –57.09ms, 95% CI[–82.79, –31.38], β = –0.09,

$F(1, 16879) = 18.95$, $p < .001$, $R^2 < .01$. When participants were categorizing targets by race, however, this tendency attenuated: $M_{\mathrm{diff}} = 0.01$ms, 95% CI[–25.69, 25.70], $\beta < 0.01$, F(1, 16899) < 0.01, $p = .99$, $R^2 < .01$ (see Figure S5).