Journal of Experimental Social Psychology 85 (2019) 103893

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Hard to disrupt: Categorization and enumeration by gender and race from mixed displays☆

Xin Yang*, Yarrow Dunham

Yale University, 2 Hillhouse Ave., New Haven, CT 06511, USA

ARTICLE INFO

Keywords:
Gender and race categorization
Face perception
Ensemble perception
Inversion

ABSTRACT

While much research has focused on the ways in which stereotyping and prejudice follow from category-based perception of others, less work has examined how and when category-based perception emerges in the first place. Here we adopt a number estimation task to explore perceivers' ability to estimate the number of individuals belonging to a given social category (race or gender) from briefly presented arrays of faces. We also investigate facial features that are crucial for this ability. Across 6 pre-registered studies (n = 461) we present novel evidence that 1) people can extract gender and race from brief displays of up to 14 faces (performance for race is better than gender estimation, approaching that in dot estimation); 2) only the encoding of gender is affected when hair is removed from faces, while the encoding of both gender and race is disrupted when luminance levels of the faces are equalized; and 3) this ability is disrupted by inversion when hair has been removed and luminance has been controlled but not when the stimuli are uncontrolled. This research investigates people's social category enumeration abilities with manipulations and nonsocial stimuli comparisons, and has implications for person perception and prejudice formation.

Consider walking through Times Square on a crowded afternoon. A sea of faces confronts you. What information about the people you encounter can you extract from a quick glance? The current research focuses on one part of this question, specifically the ability to categorize and enumerate the racial or gender makeup of an array of faces. In particular, can observers estimate the number of faces belonging to a subgroup (i.e., males, females, Black people, or White people) from arrays containing faces belonging to both race or gender categories? If so, what aspects of faces (e.g., hair type, skin tone) does this ability depend on?

Certainly humans are highly expert at face perception. Human observers form first impressions of faces in a very short period of time (e.g., as short as 33 ms) on various dimensions including competence, trustworthiness, and attractiveness. Importantly, their judgments are similar to those that develop after more exposure time, suggesting that lengthy observation is not necessary for reliable person perception (Willis & Todorov, 2006). More relevant to the present research, perceivers can quickly categorize a face based on gender and race (Freeman, Pauker, Apfelbaum, & Ambady, 2010; Martin et al., 2015; Zarate & Smith, 1990). In those studies, participants are usually shown one face at a time, and are asked to indicate its gender/race (or respond if the accompanying description about gender/race is correct or incorrect). While this research provides useful evidence that people can rapidly categorize one face based on gender and race, it does not speak to *crowd* perception, that is the ability to simultaneously categorize and enumerate *a group of* gendered and racial faces.

Perhaps the most relevant literature for crowd perception is research on ensemble perception. Ensemble perception, sometimes referred to as summary representation, is the idea that rather than faithfully representing all details of the world, our brain exploits the statistical regularities to create a summary, or "gist" of the scene, thus making perception more efficient (Ariely, 2001; Haberman & Whitney, 2011; Whitney & Yamanashi Leib, 2018). People have the visual mechanism to extract summary statistical information from visual scenes, including emotions, diversity, and hierarchy from faces (e.g., Haberman & Whitney, 2009; Phillips, Slepian, & Hughes, 2018). For gender categorization, people can rapidly perceive the sex ratio of a mixed-sex display, and this ratio further affects judgments of threat (Alt, Goodale, Lick, & Johnson, 2017) and social attitudes (Goodale, Alt, Lick, & Johnson, 2018), as well as perceiver's sense of belonging (Goodale et al., 2018). For race categorization, people can perceive the average race (e.g., Jung, Bülthoff, & Armann, 2017) and estimate the majority race (Thornton et al., 2019) from arrays of faces, and perceive difference in average emotions between two racial subgroups in a mixed-race

☆ This paper has been recommended for acceptance by Rachael Jack.

* Corresponding author at: Department of Psychology, Yale University, 2 Hillhouse Ave., New Haven, CT 06511, USA.
E-mail addresses: xin.yang@yale.edu (X. Yang), yarrow.dunham@yale.edu (Y. Dunham).

https://doi.org/10.1016/j.jesp.2019.103893
Received 25 March 2019; Received in revised form 8 August 2019; Accepted 3 September 2019

display (Lamer, Sweeny, Dyer, & Weisbuch, 2018), implying that they encode the racial identity of the constituent faces. Further, seeing emotionally segregated interracial crowds for merely 1/3 s leads to fewer biracial judgments and more racial essentialism (Lamer et al., 2018). Together, these studies demonstrate that people can extract gender and race ratios in an array of faces, with downstream influences on attitudes and judgments. However, little work has systematically explored the sensitivity or limits of this ability or how any such sensitivity relates to the ability to enumerate simpler non-social stimuli such as dot arrays, which might be thought to represent the underlying cognitive competence driving performance in such tasks (perhaps the closest exception is ; Thornton et al., 2019) they compare the ability to estimate racial compositions of crowds with more versus less controlled stimuli).

In the present work we systematically explore people's ability to enumerate the number of gendered/racial faces in a crowd. In so doing we adapt number estimation tasks used in vision science and numerical cognition literatures, where studies usually involve simple, low-level, and nonsocial stimuli like colored dots (Horne & Turnbull, 1977). In a typical dot estimation task, participants are shown brief displays of a number of dots, and are asked to indicate the total number of dots. In more complex cases there might be dots of several different colors, and participants are asked to indicate the number of dots of a specific color (i.e., they are asked to estimate a subset). This work demonstrates that perceivers can enumerate two subsets in parallel (Halberda, Sires, & Feigenson, 2006) and they appear to do so automatically and unintentionally (Cordes, Goldstein, & Heller, 2014).

In our adapted number estimation task, participants are asked to extract the numerosity of a subset (Studies 1 and 2 and replications of Studies 1 and 2), e.g., the number of female faces in a display of male and female faces. Because the locations of faces, the total number of faces, which subgroup is being asked about, and the number of faces in the subgroup all change from trial to trial, participants face the daunting task of rapidly extracting category information and estimating the numerosity of the relevant subgroup. This reflects a more difficult test than past work has included (past studies fix the total number and locations of faces, and/or ask for a majority estimate instead of numerosity estimate; see Alt et al., 2017; Goodale et al., 2018; Lamer et al., 2018; Thornton et al., 2019), but also one that more closely maps onto complex and dynamic real scenes. After establishing participants' basic ability to extract race and gender numerosities, we go on to examine the specific features that enable this categorization/enumeration ability (or put differently, what manipulations would disrupt categorization/enumeration). Inversion, hair-removal and luminance-control are common practices in this literature to examine processes of face perception. More specifically, in the literature of face perception, inversion is generally used as a way to distinguish configural vs. featural processing of faces (Leder & Bruce, 2000; Maurer, Grand, & Mondloch, 2002; Taubert, Apthorp, Aagten-Murphy, & Alais, 2011; Yin, 1969), and hair and luminance are viewed as cues for gender and race encoding both in literature and in lay beliefs. In fact, many studies begin with manipulated faces that have gone through hair removal and luminance control to explore face perception abilities in the absence of these cues (e.g., Haberman & Whitney, 2009, 2011; see Thornton et al., 2019 for an exception). In order to investigate how much each of these features (inversion, hair, and luminance) contribute to categorization of gender and race, we go on to invert faces (Studies 3), remove hair and ears from upright faces (Study 4), further grey-scale and luminosity-match faces (Study 5), and finally invert the hair-removed, luminance-matched faces (Study 6). In what follows, we describe our methodological and analytical approaches and then describe each study in detail. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in these studies.

1. Methodological approach

1.1. Stimuli

We used male and female faces of White and Black individuals from the Chicago Face Database (CFD) (Ma, Correll, & Wittenbrink, 2015) stimuli are accessible upon request at https://chicagofaces.org/default/download/). We only included faces with high category consensus so that the stimuli were unambiguously male/female or Black/White (95% agreement by CFD raters on male/female and Black/White category judgments). To include a wide range of category typicality, in each of the four sub-categories (i.e., White male, White female, Black male, and Black female), we selected 20 faces, half of them being high on both gender-typicality (masculinity/femininity) and race-typicality (Eurocentricity/Afrocentricity), and half being low on both dimensions as divided by median ratings, constituting 80 faces in the stimuli set (modified stimuli can be shared upon request; details for the specific CFD stimuli we used can be found at https://osf.io/5ka94/?view_only=15c365b391ed4e4d9d38c811ea359392). The original face images are head-to-shoulder photos, including hair and shoulder part of a gray-colored T-shirt. Each image was set to a standardized size of 213 × 150 pixels for use in the current studies and was presented against a white background.

1.2. Apparatus

All tasks were programmed using Inquisit Millisecond software package 4 (Version 4.0.10.0 (2666)) or 5 (Version 5.0.9.0 (4084)), https://www.millisecond.com/download/. For in-lab testing (Studies 1 and 2), the tasks were presented on a 15.4" MacBook Pro Laptop computer (359-mm × 247-mm). Viewing distance was unconstrained but was approximately 50 cm. For online testing (replication of Studies 1 and 2, and Studies 3–6), tasks were administered using Inquisit Web Player (Version 4.0.10 or Version 5.0.11.0 (4157)). Participants used their own computers and so testing condition was unconstrained. Approximately, each face stimulus subtended 1.15 × 1.38 degrees of visual angle.

1.3. Task design

Across studies, we adopted a number estimation task to investigate people's ability (i.e., sensitivity) to quickly categorize and enumerate gendered and racial faces from a group of faces. To do so, we presented participants multiple trials consisting of 8, 10, 12, or 14 randomly selected faces of contrasting gender (in gender estimation) or race (in race estimation). In each trial (see Fig. 1 for schematic of trials), participants first saw a fixation cross in the center of the screen (1.25 s), followed by a blank screen (1 s). Then a group of faces briefly appeared at randomized locations on the screen (1 s). After the display all faces disappeared and one randomly-selected question appeared, asking participants to indicate either how many male, female, or total faces (in gender estimation) or how many White, Black, or total faces (in race estimation) they had seen on the previous screen. Participants entered the number using keyboard and then clicked a button at the bottom of the screen to proceed to the next trial. In gender estimation trials, all faces were of the same race, but both genders appeared, with 2–11 faces being male or female[1]. Similarly, in race estimation trials, all faces were of the same gender, but both races appeared, with 2–11 faces being Black or White.

[1] Specifically, for a total set of 8 faces, numbers of faces in each of the two sub-categories could be 2, 4, or 6; for a total set of 10 faces, those numbers could be 2, 5, or 8; for a total set of 12 faces, those numbers could be 2, 4, 6, 8, or 10; for a total set of 14 faces, those numbers could be 3, 5, 7, 9, or 11. Overall, numbers of faces in a subset ranged inclusively from 2 to 11.

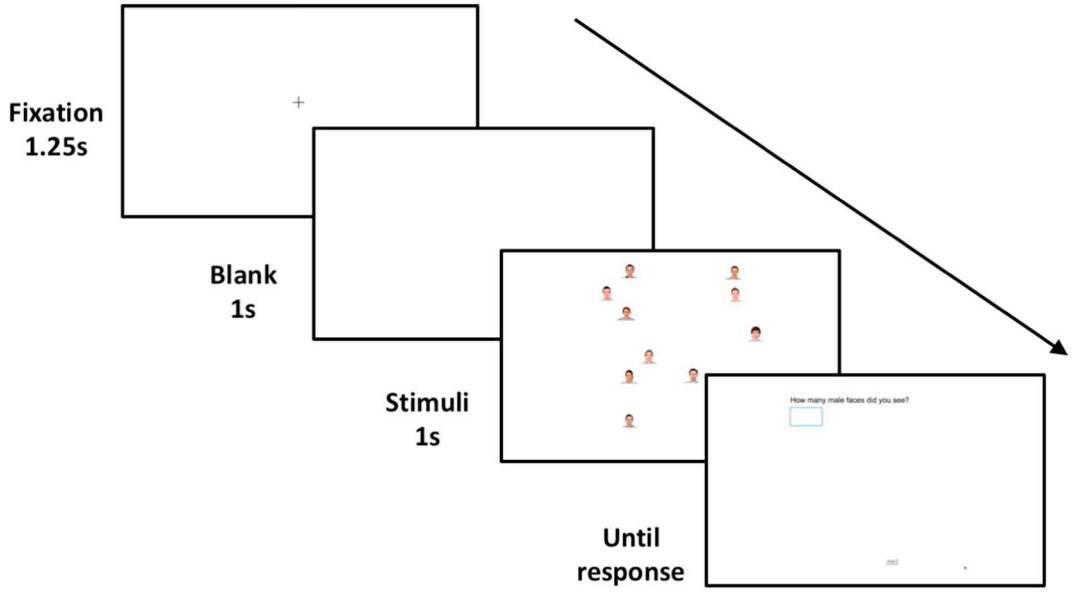

Fig. 1. Schematic of a typical number estimation (face) trial throughout all studies.

At the beginning of the experiment, participants did a practice block which consisted of 9 trials of dot estimation (i.e., estimating the number of red/blue/total dots on the screen; the schematic of trials was similar to face trials). Then they completed two experimental blocks of 45 face trials (90 trials in total), of which 60 trials were considered critical trials (asking for the number of male or female faces in gender estimation, or asking for the number of Black or White faces in race estimation). The remaining 30 trials asked for the total number of faces in the display (rather than the number of faces of a specific category). Because these trials did not require extracting category information they were broadly similar to a simple number estimation task and were included to ensure that reasonable levels of estimation sensitivity were possible on our task, i.e., with stimuli as rich as photographs of faces.

In all studies, in the end of the experiment, participants' strategies to complete the task and their demographic information were collected. To probe ability to introspect on their performance, in some of the studies (Studies 1–3), we also asked participants to give self-reports on their performance before and/or after the face trials (e.g., "how accurate do you think you will be/you were").

1.4. Analytical approach

For ease of description, we always refer to the number of target faces being asked for (i.e., the number of Black, White, male or female faces; mean-centered prior to analysis) as "target number", and the total number of faces on the screen (i.e., summing across all social categories; mean-centered) as "total faces". "Target type" refers to whether the trial was gender estimation or race estimation (dummy-coded, with gender as reference) and "question" refers to whether participants were asked about male or female faces (in gender estimation) or Black or White faces (in race estimation). In each study, before data analyses, and in accordance with our pre-registered analysis plan (see Results section), we first excluded data from participants whose mean responses were > 2 *SDs* away from the mean of all participants, which led to the elimination of data from between one and four participants per study. For included participants we also excluded any individual trial in which the participant's estimate was $+/- 3$ *SD*s away from the pooled estimate for that trial type (defined as question × target number), which constituted about 1% of the data (ranging from 1.03% to 1.26% across studies). We used linear mixed effects models, which allow the delineation of random versus fixed effects, with a random intercept for participant to respect the nested structure of the data. Data and analysis code to replicate all analyses and to create all plots can be found at: https://osf.io/5ka94/?view_only=15c365b391ed4e4d9d38c811ea359392.

For primary analyses we followed our pre-registered analysis plan for each study, which generally involved exploring the effect of one manipulated factor (e.g., inversion) related to stimulus display properties (i.e., target number and total faces). We used the "lme4" and "lmerTest" packages in R for linear mixed effects models (using the lmer() function). Our main interest was the effect of target number, i.e., how sensitive the participants were to the increase of target number. We explore sensitivity in two ways. First, as a raw measure of sensitivity, we used unstandardized regression coefficients beta (*B*). Positive betas whose confidence intervals did not include 0 are taken to indicate that participants were sensitive to the increase of target number and adjusted their answers accordingly. Because we used unstandardized betas, a participant with perfect target sensitivity would have a beta of 1, i.e., perfect encoding of every unit increase in target number. Thus, a beta of 0.4 meant that for each unit increase in target number the participant increased her estimate by 0.4 units on average (but note that unstandardized betas are sensitive to measurement scale and should not be taken as estimates of effect sizes). We also report the Pseudo-R^2 (variance explained), an effect size measure for linear mixed-effects models (following Nakagawa & Schielzeth, 2013), as the other indicator of sensitivity. This statistic is not sensitive to measurement scale and so better reflects the predict power of that term.

One potentially confounding strategy that participants could adopt would be to simply increase their estimate based on the total number of faces they perceived, without actually attending to the number of individuals from the requested sub-category. Because target number and total number are correlated, this could create the illusion of good performance. To address this all models controlled for the total number of faces in the display, but our primary variable of interest remains target number. In each of the studies, below, we report the regression models, model output, and coefficient plots for main variables (i.e., target number, total faces, and any manipulated factor). In lab Studies 1 and 2 and their online replications we observed a few effects of question (i.e., whether male, female, Black, or White targets were enumerated), but because such effects were always unpredicted, small, and inconsistent (in that they never occurred in both the initial and replication studies), we reserve discussion of them to supplemental materials.

To make direct comparisons across studies and to investigate changes in performance after each manipulation, we took two

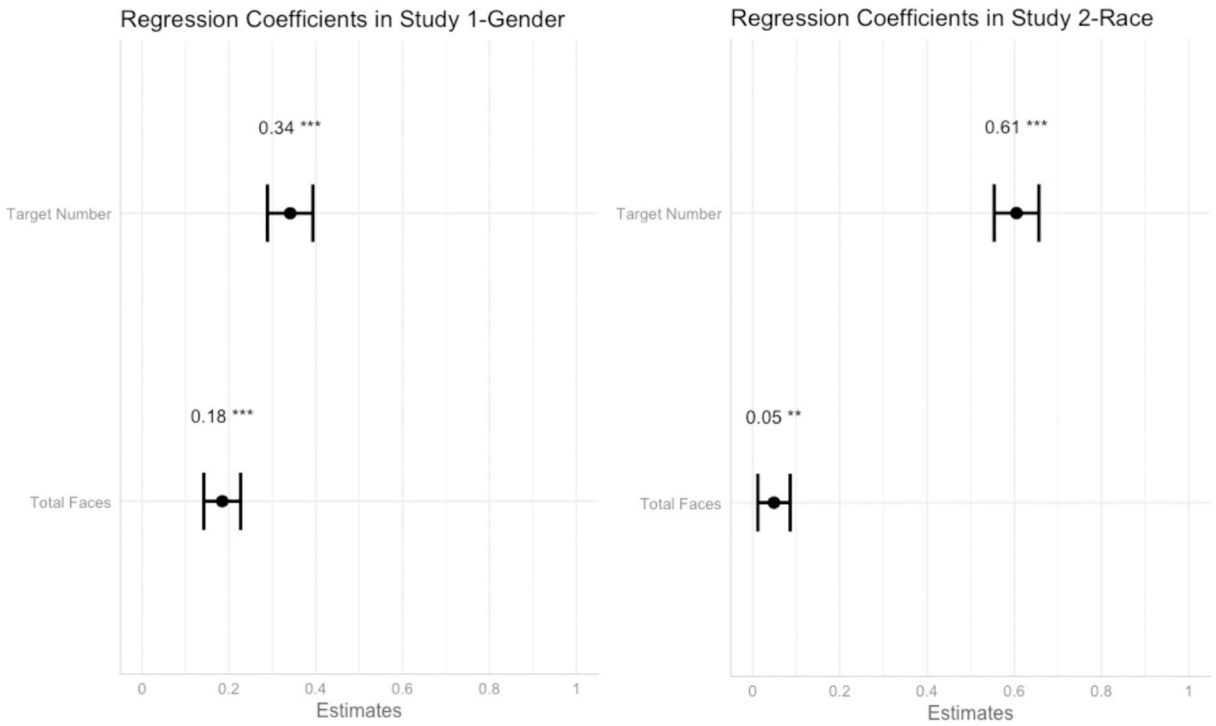

Fig. 2. Regression coefficients (on the x-axis, unstandardized regression coefficient betas and 95% confidence intervals; a beta of x meant that for each unit increase in target number the participant increased her estimate by x unit) for target number and total faces in Study 1 (Gender) and Study 2 (Race). The other predicators, question and the target number by question interaction, were also included in the model but for clarity are omitted here (see supplemental materials for the figure with all predictors).

additional analysis approach. First, we fit a set of additional models that could be consistently employed across all studies (separately for gender and race). In these models we only included target number and total faces, the two factors consistent across all studies. This allows us to compare regression coefficients for target number and for total faces across studies, as well as to compare the proportion of residual variance explained by each term across the different manipulations (as a measure of effect size), providing a complementary picture of estimation ability across all studies via a readily interpreted effect size metric. Second, to compare accuracy across studies, we calculated absolute accuracy/error scores by computing the root of the mean square of the difference between the participant's estimate and the target number (i.e., adding up the square of each estimate-target number pair, divided by the number of pairs, and computing a square root of this value). Before this calculation, we excluded any individual square error terms that was +/− 3 *SD*s away from the pooled square error terms, as these were unlikely to reflect true task compliance and reflected our pre-specified trial-level exclusion criterion.

We also asked participants for their strategies in performing the task across all studies, as well as self-reports on their own performance in Studies 1 to 3. Because we did not find their responses particularly illuminating, we present and discuss them in supplemental materials.

2. Part 1: Automatic encoding of a group of gendered and racial faces (Studies 1–2)

2.1. Study 1 and 2 (In-lab)

The aim of Study 1 and 2 was to validate our method to investigate people's ability in estimating the number of gendered (Study 1) and racial (Study 2) faces when 8 to 14 faces were briefly presented. We hypothesized that people can rapidly extract gender and race from brief displays of multiple mixed-gender and mixed-race faces (pre-registration link: http://aspredicted.org/blind.php?x=nu24a8).

2.1.1. Method

2.1.1.1. Participants. Participants were recruited from an undergraduate participant pool and received partial course credit for an introductory psychology class. We ran a pilot study to validate the study design, but since we had only limited pilot data to determine necessary power and we anticipated running another pre-registered replication study online, Study 1 also gave us an opportunity to evaluate the necessary sample size for the following studies. We determined $n = 30$ as our sample size per study following a recent related study that involves estimation of subsets and total number (Cordes et al., 2014). The final sample included 29 participants in Study 1 (13 female, $M_{age} = 18.86$, $SD_{age} = 0.80$, 55% White/European-American, 21% Asian, 10% Latino/Hispanic, 10% Mixed or Multiracial, 3% Black/African-American) and 28 participants in Study 2 (16 female, $M_{age} = 18.71$, $SD_{age} = 0.98$, 46% White/European-American, 29% Asian, 11% Mixed or Multiracial, 7% Latino/Hispanic, 7% Black/African-American). Three additional participants were also tested but were excluded from data analyses because their mean responses were 2 *SD*s away from the mean response of all participants (one in Study 1 and two in Study 2), following our pre-registered exclusion criteria. Participants were randomly assigned to complete Study 1 or Study 2.

2.1.1.2. Design. In Study 1, participants did 2 experimental blocks of gender estimation, with one all-White block containing only White faces and one all-Black block containing only Black faces (mixed-gender displays; order of blocks was randomized). Similarly, in Study 2, participants did 2 experimental blocks of race estimation, with one all-male block containing only male faces and one all-female block containing only female faces (mixed-race displays; order of blocks was randomized). The rest of the design was detailed in Methodological Approach.

2.1.2. Results and discussion

Following our pre-registered analysis plan, we excluded any estimate that is +/− 3 *SD*s away from the pooled estimate for that trial type, which constituted 1.26% and 1.03% of the data respectively. Then we used linear mixed effects models predicting estimation as a function of target number, question (i.e., questions on the number of male or female faces in gender estimation, or the number of White or Black faces in race estimation; contrast-coded), and their interaction, controlling for total faces, with a random intercept and a random slope (target number × question) for participant (for a detailed description of these terms, see Analytical Approach; see Supplemental Materials for R code).

As shown in Fig. 2, we found a significant positive effect of target number in both studies (Study 1 gender, $B = 0.34$, $SE = 0.03$, t (32.08) = 12.76, $p < .001$, partial $R^2 = 0.17$ (0.14, 0.20); Study 2 race, $B = 0.61$, $SE = 0.03$, t (29.20) = 23.12, $p < .001$, partial $R^2 = 0.47$ (0.44, 0.50)), indicating sensitivity to the increase of target number. Interestingly, in both studies, while there was a small effect of total faces (Study 1 gender, $B = 0.18$, $SE = 0.02$, t (1648.94) = 8.53, $p < .001$, partial $R^2 = 0.03$ (0.02, 0.05); Study 2 race, $B = 0.05$, $SE = 0.02$, t (1595.85) = 2.58, $p < .01$, partial $R^2 = 0.003$ (0.00, 0.01)), the inclusion of this term did not eliminate or even much reduce the effect of target number, suggesting that estimation was largely driven by sensitivity to the target category rather than inferred from the total number of faces in the display (see supplemental materials for additional effects involving question).

Consistent with our predictions, we found that people can extract gender and race from brief displays of a group of gendered and racial faces, and that their estimation was due to target sensitivity (not merely adjusting for total faces). Further, participants were considerably more sensitive in the race task than in the gender task, as clearly visible in Fig. 2. We also provide Figs. 3 and 4 to show the effect of target sensitivity more clearly (for regression plots for the following studies, see supplemental materials). We note that visual inspection of these regression plots suggests a "compression effect" in which participants tended to over-estimate small subset sizes and under-estimate large subset sizes. While speculative, this might have occurred because some participants attempted to estimate the number of a subset by dividing their estimate of the total number by 2 (e.g., if there were 12 faces in total, estimating 6 of each subset). This strategy could create this compression effect, though it is important to note that it cannot explain our larger pattern of results because such a strategy would entail that only total faces, and not target number, predicts responses.

A simulation-based power analysis using Study 1 data (using the "simr" package in R) showed that we had sufficient power (> 90%) to detect a target number effect (i.e., regression coefficient beta) as small as 0.1, an effect below which we would consider estimation ability to be inconsequential, and an effect < 1/3rd what we observed in Study 1. Therefore, our studies were well-powered with $n = 29$ sample size. However, because we anticipated the possibility of increased noise and data loss in the online replication, due, for example, to inattentive participants, we doubled the sample size to $n = 60$.

2.2. Replication of Study 1 and 2 (MTurk)

We sought to replicate the results of Study 1 and 2 using an online sample on Amazon Mechanical Turk (MTurk), with the goal of running future studies online should we be able to replicate critical effects. The final sample included 64 participants in Study 1 replication (31 female, $M_{age} = 39.94$, $SD_{age} = 12.30$, 86% White/European-American, 6% Black/African-American, 3% Asian, 3% Latino/Hispanic, 2% Other) and 62 participants in Study 2 replication (35 female, $M_{age} = 36.97$, $SD_{age} = 9.48$, 79% White/European-American, 10% Black/African-American, 6% Asian, 3% Latino/Hispanic, 2% Mixed or Multiracial), after excluding 7 participants (3 in Study 1 replication and 4 in Study 2 replication) because their mean responses were > 2 *SDs* away from the mean of all participants. Participants were randomly assigned to complete one of the two studies. We analyzed data following our pre-registered analysis plan (pre-registration link: http://aspredicted.org/blind.php?x=yi56xx) and as shown in Fig. 5, we replicated the main findings in Study 1 and 2 (for the effect of target number, in Study 1 replication $B = 0.34$, $SE = 0.03$, t (67.12) = 13.38, $p < .001$, partial $R^2 = 0.14$ (0.12, 0.16); in Study 2 replication $B = 0.64$, $SE = 0.03$, t (63.84) = 19.14, $p < .001$, partial $R^2 = 0.34$ (0.32, 0.37)) (see supplemental materials for additional effects involving question).

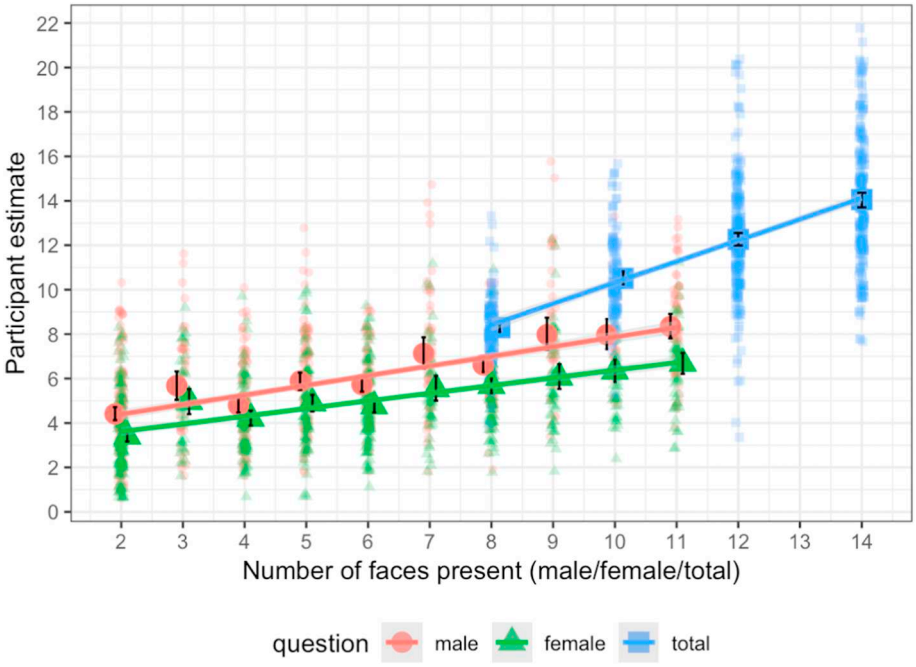

Fig. 3. Regression plot for male, female, and total faces in Study 1 (Gender). Points represent individual estimates (jittered); lines reflect linear regression predicting estimate from target number.

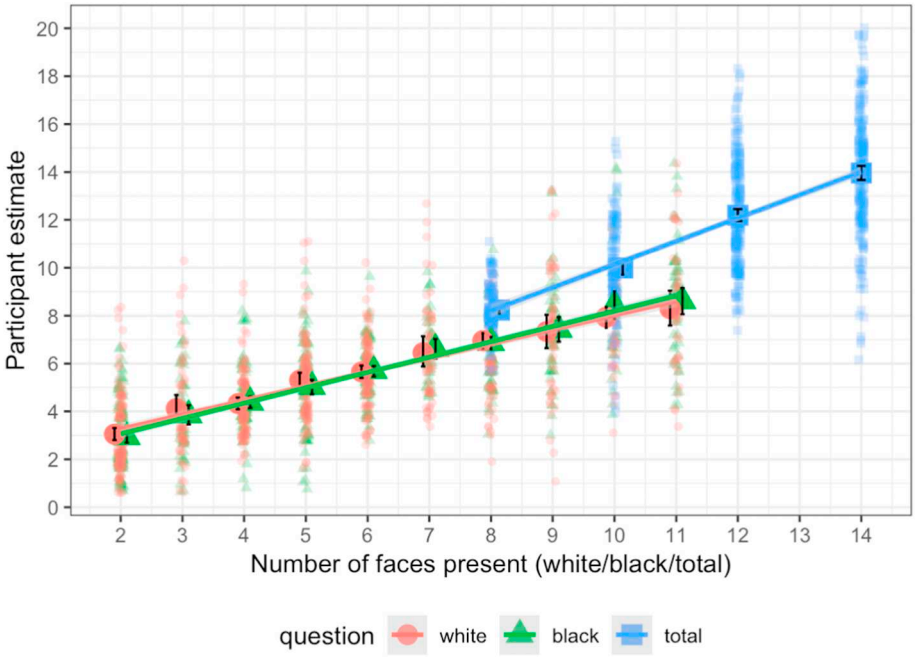

Fig. 4. Regression plot for White, Black, and total faces in Study 2 (Race). Points represent individual estimates (jittered); lines reflect linear regression predicting estimate from target number.

3. Part 2: Manipulations to disrupt categorization and enumeration (Studies 3–6; Online)

Having successfully replicated the main results of Study 1 and 2 using online samples, we continued the investigation with online participants for Studies 3 through 6 (Studies 3–5: tested on MTurk; Study 6: tested an undergraduate participant pool online). Our goal was to gradually modify the faces and thereby explore what aspects of the faces were driving performance in the first studies. In Study 3, we used both upright and inverted faces to explore the effect of inversion on social category enumeration. In Study 4, we manually removed hair (and ears) from faces to explore the role of these features in enumeration. In Study 5, we equalized the mean luminance of the hair-removed faces using the SHINE toolbox (Willenbockel et al., 2010) in

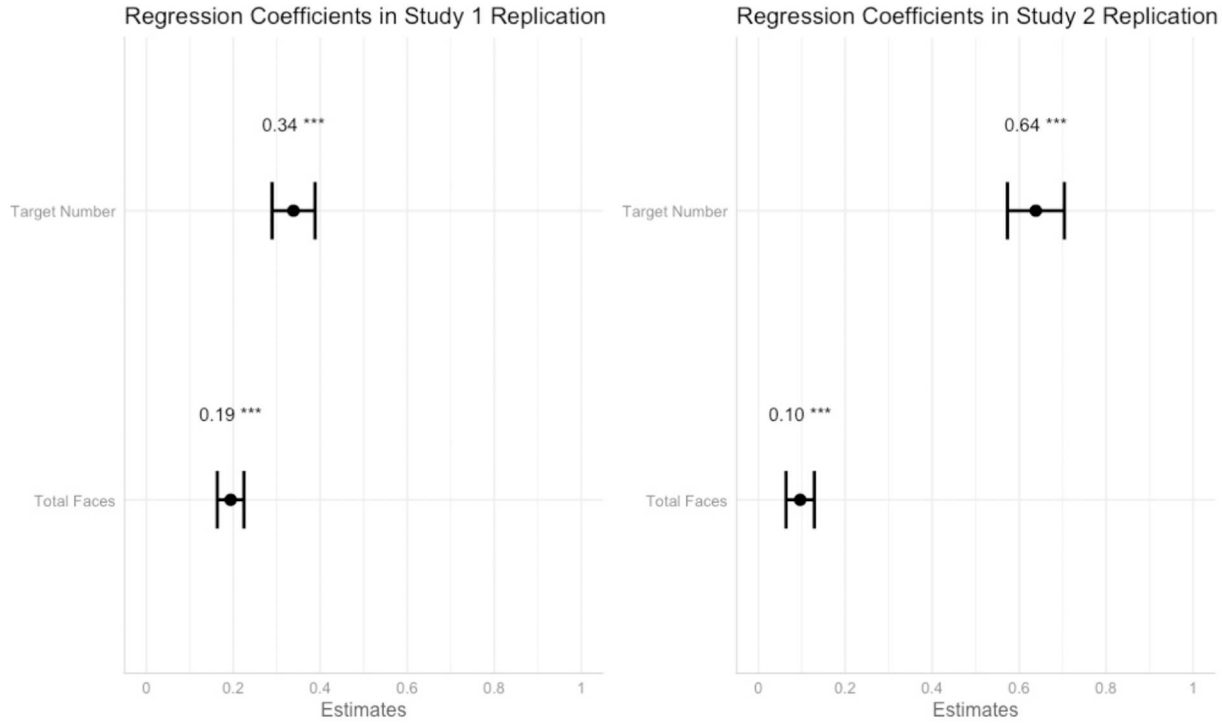

Fig. 5. Regression coefficients (on the x-axis, unstandardized regression coefficient betas and 95% confidence intervals; a beta of x meant that for each unit increase in target number the participant increased her estimate by x unit) for target number and total faces in Replication Study 1 (Gender) and Replication Study 2 (Race). The other predicators, question and the target number by question interaction, were also included in the model but for clarity are omitted here (see supplemental materials for the figure with all predictors).

Matlab, allowing us to explore the role of luminosity in driving performance. Finally, in Study 6, we inverted the hair-removed luminance-controlled faces to explore the combined effects of all these manipulations on enumeration. We predicted that all or most of these manipulations would negatively affect performance, though it was an open question what manipulations would have the largest effects or if any of these manipulations would completely disrupt encoding, as well as if degradations in performance would occur similarly for race and gender. In an effort to simplify the design from prior studies, we only included White faces in gender trials and male faces in race trials.

3.1. Study 3: inverting faces

3.1.1. Method

3.1.1.1. Participants and design. In Study 3 (pre-registration link: http://aspredicted.org/blind.php?x=mw9ek9), the final sample included 58 participants (25 female, M_{age} = 35.86, SD_{age} = 11.24, 78% White/European-American, 7% Asian, 7% Black/African-American, 7% Latino/Hispanic, 2% Mixed or Multiracial). They were randomly assigned to gender estimation version (n = 27) or race estimation version (n = 31). Participants completed both an upright face block and an inverted face block in a randomized order. An additional 4 participants (3 in gender version, 1 in race version) were tested but excluded because their mean responses were > 2 *SDs* away from the mean of all participants in that version.

3.1.2. Results and discussion

Following our pre-registered analysis plan, we first excluded any trial-level outliers, which constituted 1.11% of the data in each version. Then we used two linear mixed effects models (one for gender version and one for race version) predicting estimation as a function of target number, inversion (whether the faces were inverted or upright; dummy-coded, with upright faces as reference), and their interaction, controlling for total faces, with a random intercept and a random slope (target number × inversion) for participant. As shown in Fig. 6, we found a significant effect of target number in both versions (Study 3 gender, B = 0.31, SE = 0.04, t (28.81) = 7.96, p < .001, partial R^2 = 0.07 (0.05, 0.10); Study 3 race, B = 0.60, SE = 0.05, t (32.09) = 11.39, p < .001, partial R^2 = 0.21 (0.17, 0.24)), indicating target sensitivity. Unexpectedly, there was no effect of inversion (ps > 0.71) or target number × inversion interactions (ps > 0.21); inversion did not seem to harm performance. Additionally, in both versions there was a significant effect of total faces (Study 3 gender, B = 0.29, SE = 0.02, t (1540.65) = 12.04, p < .001, partial R^2 = 0.07 (0.05, 0.10); Study 3 race, B = 0.11, SE = 0.02, t (1765.91) = 4.62, p < .001, partial R^2 = 0.01 (0.00, 0.02)), with the effect of total faces in the gender version approaching the size of the effect of target number.

Contrary to our prediction, results showed that people can extract gender and race from inverted faces as well as they do with upright ones. This finding seems at odds with the past research documenting the face inversion effects in configural or holistic face processing (i.e., perceiving relations among facial features; Freire, Lee, & Symons, 2000; Taubert et al., 2011). However, past studies also show that in featural processing of faces (i.e., processing of featural cues such as luminance, hair, and eyes), there is no face inversion effect (Freire et al., 2000). A possible explanation for our finding is that with unmanipulated faces (i.e., with hair/ears and real colors), there are enough featural cues to encode gender and race information even when faces are inverted (for a discussion of how using artificial, well-controlled faces might affect inversion results, see Valentine, 1988). We return to this in the General Discussion.

However, we noticed that total faces might play a larger role in predictive power than in previous studies because regression coefficients for total faces increased, especially in gender estimation. We also return to this issue in the General Discussion.

In the following studies, we turned to directly manipulating the features of faces. In Study 4, we removed hair (and ears) from faces, leaving only the main part of the face intact. We predicted that people's performance in extracting gender would suffer and would suffer more than extracting race, because hair is likely a stronger cue for gender than for race. We had no prediction regarding whether people's performance in extracting race would suffer or not.

3.2. Study 4: removing hair

3.2.1. Method

3.2.1.1. Participants and design. In Study 4 (pre-registration link: http://aspredicted.org/blind.php?x=eu7zw2), the final sample included 58 participants (22 female, M_{age} = 34.14, SD_{age} = 8.61, 74% White/European-American, 10% Black/African-American, 9% Asian, 5% Mixed or Multiracial, 2% Latino/Hispanic) after excluding 2 participants because their mean responses were > 2 *SDs* away from the mean of all participants. Each participant completed both gender estimation block (all White faces) and race estimation block (all male faces) in a randomized order, where all faces were hair-removed.

3.2.2. Results and discussion

Following our pre-registered analysis plan, we first excluded any trial-level outliers, which constituted 1.23% of the data. In order to examine the effect of hair removal in comparison to the original data, we combined the data (hair was removed) with the Study 1 online replication data (using the all-White block; hair was not removed) and the Study 2 online replication data (using the all-male block; hair was not removed). We started with our pre-registered linear mixed effects model, predicting estimation as a function of target number, target type, hair-removal (whether hair was removed; dummy-coded, with non-removed faces as reference), and their interactions, controlling for total faces, with a random intercept and a random slope (target number) for participant. We found a marginally significant 3-way interaction (p = .066), so we decomposed into 2 models, one each for gender and for race (see Fig. 7 for results). In gender estimation, there were significant effects of target number (B = 0.28, SE = 0.03, t (136.26) = 10.97, p < .001, partial R^2 = 0.04 (0.03, 0.06)) and total faces (B = 0.32, SE = 0.02, t (3473.89) = 16.93, p < .001, partial R^2 = 0.06 (0.05, 0.08)), with more predictive power from total faces than from target number, though target number remained similar in magnitude to prior studies. We also found a significant effect of hair-removal (B = 0.91, SE = 0.19, t (119.45) = 4.69, p < .001, partial R^2 = 0.03 (0.02, 0.04)), but the interaction between hair-removal and target number was not significant (p > .82). Thus, participants made larger estimates for hair-removed faces compared to unmanipulated ones but the acuity of their estimations was not affected. In race estimation, there were significant effects of target number (B = 0.70, SE = 0.04, t (121.49) = 19.26, p < .001, partial R^2 = 0.26 (0.24, 0.29)) and total faces (B = 0.08, SE = 0.02, t (3378.54) = 5.00, p < .001, partial R^2 = 0.01 (0.00, 0.01)). There was also a marginally significant effect of the interaction between hair-removal and target number (B = 0.09, SE = 0.05, t (116.24) = 1.76, p = .08, partial R^2 = 0.002 (0.00, 0.01)), suggesting that removing hair slightly *improved* performance, albeit only marginally.

Next, in Study 5 we would further equalize the mean luminance of the faces to explore whether we could disrupt the encoding of race (and maybe further disrupt the encoding of gender).

3.3. Study 5: equalizing luminance

3.3.1. Method

3.3.1.1. Participants and design. In Study 5 (pre-registration link: http://aspredicted.org/blind.php?x=6iz28u), the final sample included 58 participants (27 female, M_{age} = 37.41, SD_{age} = 10.48, 76% White/European-American, 9% Asian, 9% Black/African-American, 3% Latino/Hispanic, 2% Mixed or Multiracial, 2% Prefer

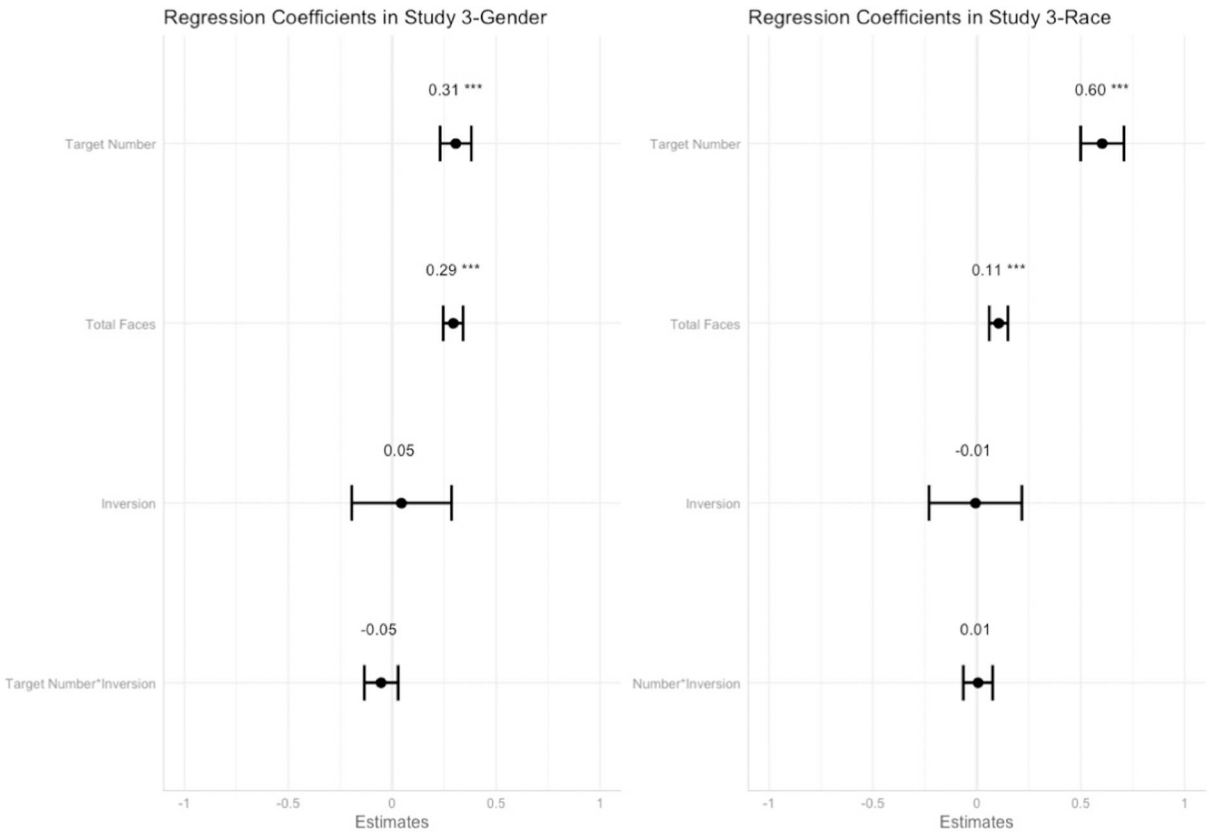

Fig. 6. Regression coefficients (and 95% confidence intervals, unstandardized) for target number, total faces, inversion, and target number × inversion interactions in Study 3-Gender and Study 3-Race (upright faces as reference). Results showed that inversion did not harm performance.

not to say) after excluding 2 participants because their mean responses were > 2 *SDs* away from the mean of all participants. They completed both gender estimation block (all White faces) and race estimation block (all male faces) where all faces were hair-removed *and* luminance-controlled.

3.3.2. Results and discussion

Following our pre-registered analysis plan, we first excluded trial-

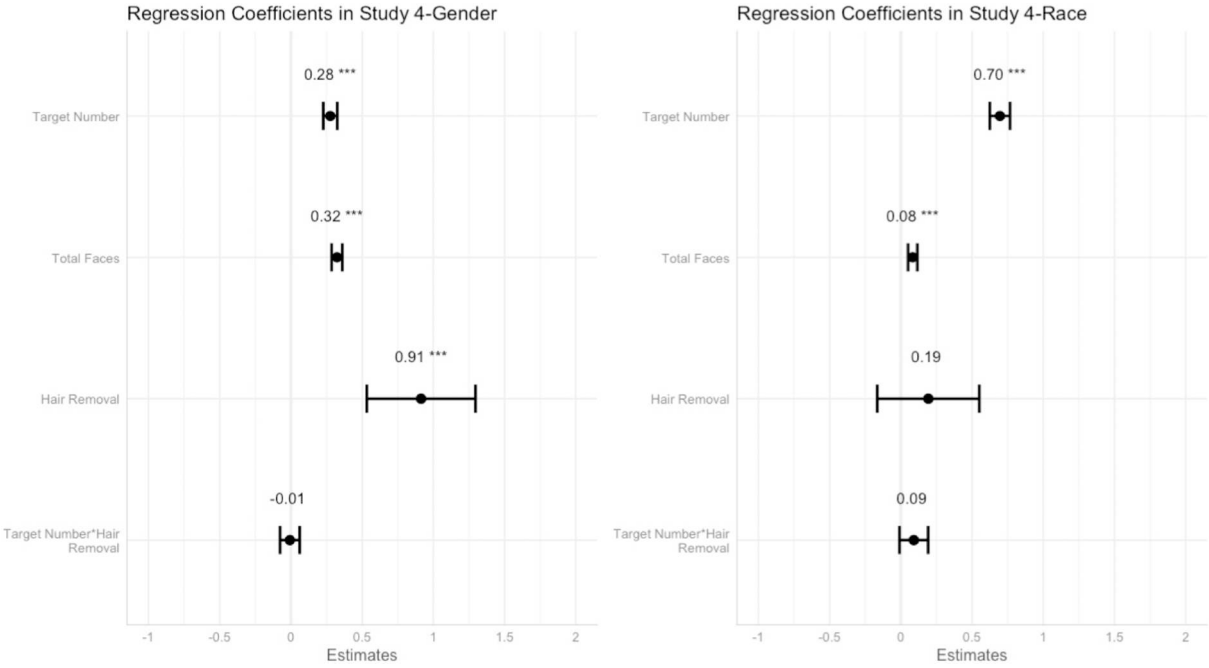

Fig. 7. Regression coefficients (and 95% confidence intervals, unstandardized) for target number, total faces, hair removal, and target number × hair removal interactions in Study 4-Gender (combining with Study 1 replication data) and Study 4-Race (combining with Study 2 replication data) (with non-removed faces as reference). Results showed that hair removal affected gender estimation (by leading to larger estimates) but not race estimation.

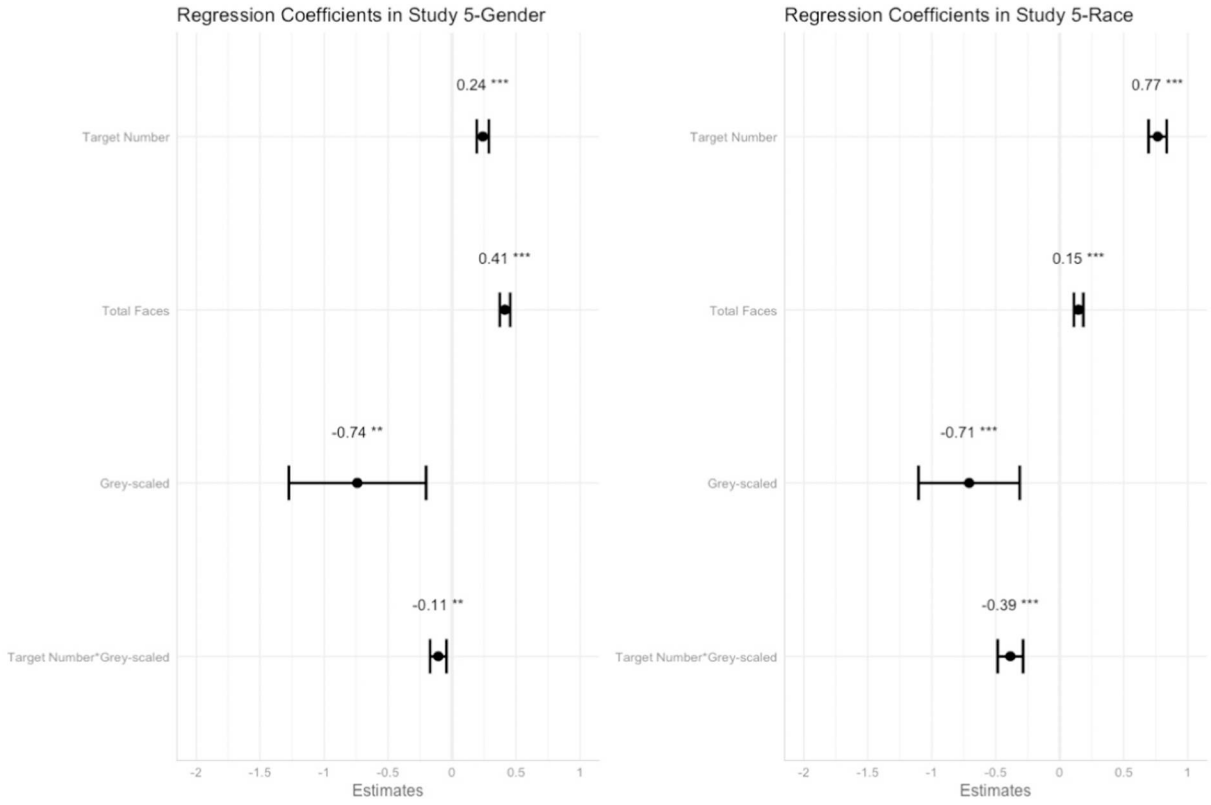

Fig. 8. Regression coefficients (and 95% confidence intervals, unstandardized) for target number, total faces, grey-scaling/luminance-control, and target number × grey-scaling/luminance-control interactions in Study 5-Gender and Study 5-Race (combining with Study 4 data where faces were hair-removed but not luminance-controlled; non-luminance-controlled stimuli as reference). Results showed that equalizing luminance harmed both gender and race estimation (by decreasing target sensitivity).

level outliers, which constituted 1.05% of the data. Then we combined these data (hair-removed plus luminance-controlled faces) with Study 4 data (hair-removed faces). We started with a linear mixed effects model similar to that in Study 4, predicting estimation as a function of target number, target type, luminance-control (whether luminance was equalized; dummy-coded, with non-luminance-controlled stimuli as reference), and their interactions, controlling for total faces, with a random intercept and a random slope (target number) for participant. We found a significant 3-way interaction ($p < .001$), so we decomposed into 2 models, one each for gender and for race (see Fig. 8 for results). In gender estimation, there were significant effects of target number ($B = 0.24$, $SE = 0.02$, t (129.89) = 9.96, $p < .001$, partial $R^2 = 0.03$ (0.02, 0.04)) and total faces ($B = 0.41$, $SE = 0.02$, t (3307.78) = 19.91, $p < .001$, partial $R^2 = 0.08$ (0.07, 0.10)), indicating target sensitivity and a large effect of total faces. We also found a significant effect of luminance-control ($B = -0.74$, $SE = 0.27$, t (113.40) = −2.70, $p < .01$, partial $R^2 = 0.02$ (0.01, 0.03)) and an interaction between target number and luminance-control ($B = -0.11$, $SE = 0.03$, t (113.98) = −3.23, $p < .01$, partial $R^2 = 0.003$ (0.00, 0.01)), suggesting that participants were less sensitive to target increase after this manipulation. In race estimation, there were significant effects of target number ($B = 0.77$, $SE = 0.04$, t (116.79) = 21.23, $p < .001$, partial $R^2 = 0.27$ (0.24, 0.29)) and total faces ($B = 0.15$, $SE = 0.02$, t (3247.63) = 7.95, $p < .001$, partial $R^2 = 0.02$ (0.01, 0.02)), indicating considerable ability to extract numerosity from the arrays. Also, the effect of luminance-control ($B = -0.71$, $SE = 0.20$, t (113.54) = −3.51, $p < .001$, partial $R^2 = 0.02$ (0.01, 0.03)) and the interaction between target number and luminance-control ($B = -0.39$, $SE = 0.05$, t (110.82) = −7.67, $p < .001$, partial $R^2 = 0.05$ (0.04, 0.06)) were significant, suggesting that participants were less sensitive to target increase when luminance was equalized.

As predicted, we found that people's performance in extracting race and gender suffered when luminance was equalized, as indicated by a significant decrease in target sensitivity for both race and gender, though effects were greater for race than gender. Moreover, we observed larger regression coefficients for total faces and smaller regression coefficients for target number compared to all previous studies, suggesting that participants' estimates were influenced more by total faces than target number as the task became harder, a topic we return to in the General Discussion. Next, in Study 6 we further invert these hair-removed and luminance-controlled faces to explore the joint effects of these manipulations.

3.4. Study 6: inversion again

3.4.1. Method

3.4.1.1. Participants and design. In Study 6 (pre-registration link: http://aspredicted.org/blind.php?x=de4bw2), the final sample included 104 participants (62 female, $M_{age} = 19.33$, $SD_{age} = 1.23$, 39% White/European-American, 22% Asian, 14% Latino/Hispanic, 10% Mixed or Multiracial, 9% Black/African-American, 4% Prefer not to say, 2% Other). They were randomly assigned to gender estimation version ($n = 52$), where all faces were White, or race estimation version ($n = 52$), where all faces were male. We doubled the sample size (compared to the first inversion study, Study 3) because we aimed to increase the power to detect inversion effects. Participants completed both upright face block and inverted face block in a randomized order, where all faces were hair-removed and luminance-controlled. An additional 6 participants (3 in gender version, 3 in race version) were tested but excluded because their mean responses were > 2 *SDs* away from the mean of all participants in that version.

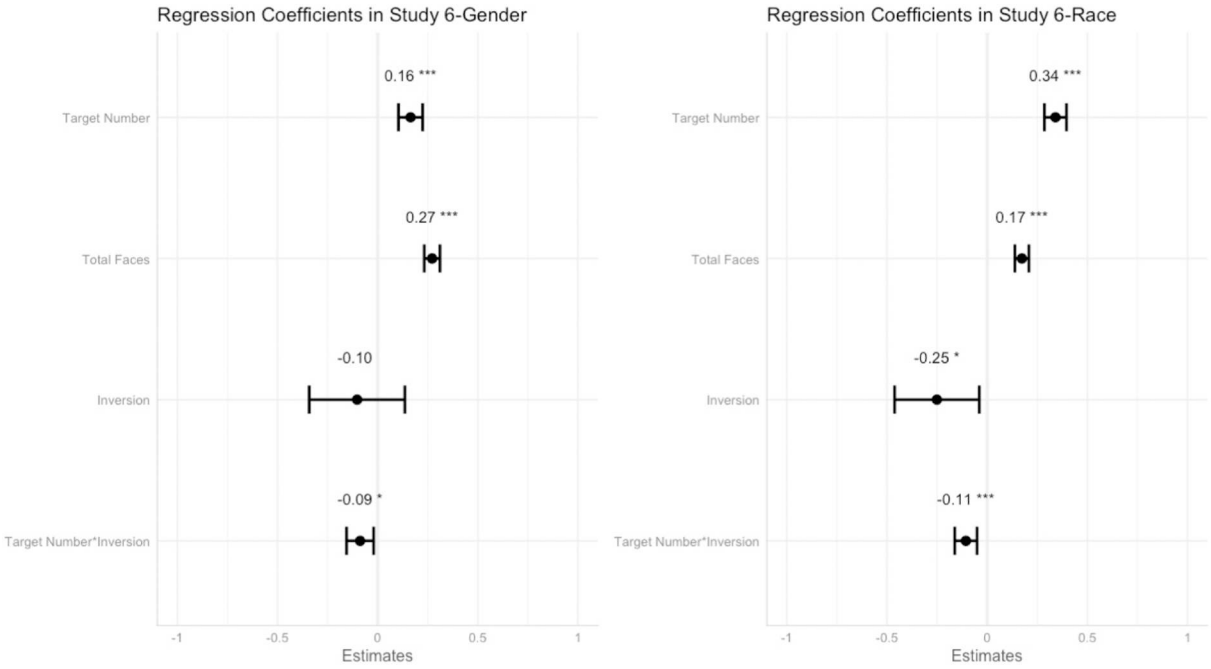

Fig. 9. Regression coefficients (and 95% confidence intervals, unstandardized) for target number, total faces, inversion, and target number × inversion interactions in Study 6-Gender and Study 6-Race (upright faces as reference). Results showed that inversion harmed both gender and race estimation (mainly by decreasing target sensitivity).

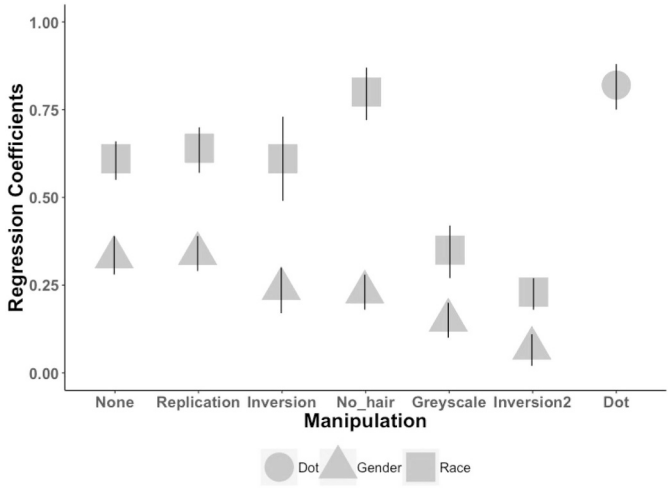

Fig. 10. Comparison of target sensitivity. This figure displays regression coefficients for target number across studies/manipulations. Error bars are 95% *CIs*. Sensitivity to target number decreased as we manipulated the faces and the task became harder. Nonetheless, in all cases target sensitivity was stronger for race (which was similar to that of dot estimation) than for gender estimation.

3.4.2. Results and discussion

Following our pre-registered analysis plan, we first excluded any trial-level outliers, which constituted 1.13% (gender) and 1.11% (race) of the data respectively. Then we used two linear mixed effects models, one for each version (employing the same models as in Study 3; see Fig. 9 for results). We found significant effects of target number (gender, $B = 0.16$, $SE = 0.03$, t (55.25) $= 5.33$, $p < .001$, partial $R^2 = 0.02$ (0.01, 0.03); race, $B = 0.34$, $SE = 0.03$, t (55.99) $= 12.12$, $p < .001$, partial $R^2 = 0.08$ (0.07, 0.10)) and total faces (gender, $B = 0.27$, $SE = 0.02$, t (2945.82) $= 13.55$, $p < .001$, partial $R^2 = 0.05$ (0.04, 0.06); race, $B = 0.17$, $SE = 0.02$, t (2978.01) $= 9.75$, $p < .001$, partial $R^2 = 0.03$ (0.02, 0.04)) in both versions. Importantly, there were significant effects of the interaction between inversion and target number (gender, $B = -0.09$, $SE = 0.03$, t (52.72) $= -2.54$, $p = .01$, partial $R^2 = 0.003$ (0.00, 0.01); race, $B = -0.11$, $SE = 0.03$, t (50.92) $= -3.76$, $p < .001$, partial $R^2 = 0.01$ (0.00, 0.01)) in both gender and race estimation; participants were less sensitive to target increase when the manipulated faces were inverted. Additionally, in race estimation, there was also a significant effect of inversion, $B = -0.25$, $SE = 0.11$, t (49.04) $= -2.33$, $p = .02$, partial $R^2 = 0.003$ (0.00, 0.01).

While we did not include it in our pre-registered analysis plan, our data also provide the opportunity to compare inversion effects between gender and race estimation. To do so, following the models in Study 4 and 5, we also fit a linear mixed effects model predicting estimation as a function of target number, target type, inversion (dummy-coded, with upright faces as reference), and their interactions, controlling for total faces, with a random intercept and a random slope (target number × inversion) for participant. We did not find any significant interactive effects involving target type and inversion ($ps > 0.39$); inversion influenced gender and race estimation to a similar extent.

Taken together, we found that participants performed worse in extracting gender and race with inverted faces but only when those faces were hair-removed and luminance-controlled. We note that this differs from Study 3 in which inverted but otherwise uncontrolled faces were used as stimuli and inversion did not affect performance. We return to this difference in the General Discussion.

4. Comparison among all studies

We fit a set of additional models separately for gender and race that could be consistently employed across all studies (see Analytical Approach). Fig. 10 shows regression coefficients for target number across manipulations, with performance in the dot estimation task (practice block data for all participants across all studies) as another comparison. Table 1 shows the error score and proportion of residual variance explained by each term across different manipulations[2].

[2] Specifically, we started with a null model predicting gender/race estimation responses with only a random intercept for participant. We then sequentially added first total faces or first target number in order to estimate the residual

Table 1
Comparison of absolute accuracy (indicated by the error scores) and explained variance across studies. As we manipulated the faces and the task became harder, error scores increased and overall explained variance decreased, with less variance explained by target number and more variance explained by total faces. In all cases, target number explained more variance (and total faces explained less variance) in race estimation compared to gender estimation (in fact, before grey-scaling the faces/luminance-control, race estimation was comparable to dot estimation), again suggesting that participants were better at race estimation than gender estimation.

Manipulation	*n*	Target type	Error	Only target	Only total	Overall explained variance
None	29	Gender	2.55	24.11%	12.35%	26.77%
None	28	Race	1.85	53.55%	9.35%	53.70%
(None) Replication	64	Gender	2.59	21.22%	10.82%	23.58%
(None) Replication	62	Race	2.31	42.43%	10.44%	42.95%
Inversion	27	Gender	2.76	18.69%	20.27%	27.24%
Inversion	31	Race	2.26	41.71%	10.71%	42.17%
No hair	58	Gender	3.31	12.59%	18.43%	22.41%
No hair	58	Race	1.99	56.68%	11.12%	56.78%
Greyscale	58	Gender	3.28	8.41%	15.27%	17.53%
Greyscale	58	Race	2.86	18.82%	11.37%	21.87%
Inversion 2	52	Gender	3.19	3.54%	9.96%	10.43%
Inversion 2	52	Race	2.74	13.71%	9.25%	16.81%
Dot	461	Dot	2.52	38.31%	25.56%	40.68%

5. General discussion

Adapting the number estimation task to the study of social categorization, we investigated the categorization and enumeration of gender and race from brief presentations of dynamic displays of multiple faces. Our findings show that people have the ability to not merely attend to, but also rapidly extract the social identities contained in crowds. This ability might well be useful when navigating complex visual scenes, such as walking through a crowded square, but it may also lead to less positive outcomes such as the perception of threat when the crowd composition trends towards negatively stereotyped categories (e.g., Alt et al., 2017). This ability also plausibly relates to the more general ability to extract meaningful summary statistics from complex scenes (e.g., Haberman & Whitney, 2011); in this case, the summary statistics are the approximate numeric magnitude of various social category constituencies.

Across 6 studies we found that people can extract gender and race from brief displays, though performance is better for race estimation than gender estimation. Importantly, performance on race estimation approached that of dot estimation, suggesting that our ability to represent and categorize race is on a par with that of simpler stimuli. We also explored the limits of this ability across various manipulations of the faces. Only the encoding of gender was affected when hair (and ears) was removed from faces but the encoding of both gender and race was disrupted when the mean luminance levels of the faces were equalized. Surprisingly, the ability to enumerate gendered and racial faces was not affected by inversion when the stimuli were uncontrolled faces; inversion only harmed performance when we used controlled faces (i.e., after hair-removal and luminance-control).

Our findings are consistent with past research on face perception that shows people's ability to perceive gender and race from faces (e.g., Alt et al., 2017; Freeman et al., 2010; Lamer et al., 2018; Thornton et al., 2019; Zarate & Smith, 1990). In our research, we further explored the sensitivity of this effect and showed the change in face perception abilities across manipulations and comparisons with simpler stimuli. We provided strong evidence that shows people's reasonably good enumeration abilities with manipulations and nonsocial stimuli comparisons: people could still categorize and enumerate gendered and racial faces even after inversion and manipulations to remove hair and control luminance, and they could enumerate racial faces at a similar level as enumerating colored dots.

One seemingly surprising finding was the lack of inversion effects in Study 3. Using full-color head-to-shoulder faces, we did not find any difference between enumerating upright faces and inverted ones. According to the literature on face inversion effects, for configural or holistic face processing, that is, when processing depends on perceiving the relations among facial features, inversion harms performance (Freire et al., 2000; Taubert et al., 2011). By contrast, for featural processing of faces (i.e., processing of featural information like luminance, hair, eye color, eye brows, chin, and face shape), there are often no face inversion effects (Freire et al., 2000). Taken together, this might suggest that gender and race categorization in our Study 3 depended on featural processing. Interestingly, there is ongoing debate on whether gender/race categorization relies on configural/holistic processing (Baudouin & Humphreys, 2006; Caharel et al., 2011; Stevenage & Osborne, 2006), on featural processing (Brown & Perrett, 1993), or either/both depending on tasks and stimuli (Dupuis-Roy, Fortin, Fiset, & Gosselin, 2009). Returning to our Study 3, one possibility is that the use of full-color photographs in this study might facilitate featural processing of faces, thereby eliminating potential effects of inversion. We note that most past work on inversion effects used controlled stimuli that lack some features, such as grey-scaled or computer-generated black-and-white faces, or drawings of faces that were also black-and-white (Valentine, 1988). Similarly, with hair-removed grey-scaled faces in our Study 6, we also found inversion effects. By reducing the salience of featural cues, such stimuli might increase reliance on configural cues and so increase the observed effects of inversion. Unraveling this tension would be a fruitful avenue for future work.

One important design in our studies was manipulating facial features to disrupt the encoding of gender and race. Our first step was to remove hair and ears from faces, a classic procedure used in the literature (in fact, many studies begin by controlling faces in this way). We expected that performance would decrease and would decrease more in gender estimation since hair is considered a salient feature aiding gender categorization. Surprisingly, estimates were not degraded when hair was removed, though perceivers did give somewhat larger estimates overall, perhaps suggesting they *thought* their performance was worse and so corrected by giving slightly higher estimates. For race estimation, we did not find any changes in performance.

The manipulation that was most important in disrupting race categorization was controlling for mean luminance. Indeed, after making faces black-and-white and also equalizing the mean luminance of White and Black faces, in Study 5, we saw a substantial decrease in performance in race estimation compared to prior studies. Also, we found that performance in gender estimation was affected, though perhaps not as dramatically. We note that by controlling for mean-luminance, we also

(*footnote continued*)
incremental variance explained by the addition of that factor. We also report overall explained variance when we included both factors.

grey-scaled the faces so that all faces became black-and-white. As real-color faces are converted to black-and-white, some facial features and gendered characteristics involving pigmentation might be lost, which might harm gender estimation. In fact, color cues are important in face recognition and higher-level vision in general (Tanaka, Weiskopf, & Williams, 2001; Yip & Sinha, 2002). More relevant to current research, color/pigmentation of faces plays a large role in gender classification (Hill, Bruce, & Akamatsu, 1995; Nestor & Tarr, 2008; Tarr, Kersten, Cheng, & Rossion, 2001). Future research aiming to tease these two effects apart could manipulate colors of faces in several steps, for example first grey-scaling faces and then controlling the mean luminance, measuring estimation abilities after each step.

One interesting aspect of our results concerns the relative magnitude of the effects of target number and total faces. In early studies the target number effect was always larger than the total faces effect, suggesting that participants were (at least primarily) directly estimating the relevant subsets (i.e., increasing their estimates according to the increase of target faces not the increase of total faces). However, in later studies, as the effect of target number declined, the effect of total faces tended to increase. By the final studies on gender the predictive power associated with total faces was larger than that of target number (see Table 1). What does this mean? We think that when stimuli were degraded, participants who did not think they were able to estimate subsets might instead employ an alternative strategy. More specifically, they might increase their estimates of the queried subset in relation to their estimate of total faces, for example by multiplying their total faces estimates, which were always highly accurate, by a rough ratio corresponding to their intuition of the subset sizes. This would reflect a strategy shift. In other words, as the task became harder, perceivers may have found their ability to extract the relevant subset compromised, and may have come to rely more on the correlated cue of total set size, assuming that larger arrays contained larger subsets. This "total faces" strategy, estimating total faces and the gender/race ratio, parallels that sometimes seen in the ensemble perception literature, where research shows that people can "compute" the mean and standard deviation of a group of stimuli by perceiving the crowd as an "ensemble" (e.g., Haberman & Whitney, 2009). Since such an "ensemble" is a rather coarse summary statistic of the visual scene, performance is somewhat worse than when participants directly estimate subsets. To further explore the possibility of strategy shifts, future studies could directly manipulate strategy by providing instructions, i.e., asking participants to focus on subsets or total faces, and compare performances.

The current research leaves many interesting open questions for future studies. First and foremost, one important issue regards levels of "automaticity" of social categorization. In our studies we did not tell participants which question would be asked prior to stimulus display, but they always knew it was about gender or race. In other words, we explicitly led participants to pay attention to social categories and then measured their ability to encode them. Therefore, with the current paradigm we cannot show whether people could or would enumerate gendered and racial faces if they did not know that the study was about gender or race. The present paradigm is not ideally suited to this question, however, because participants would only be truly naïve prior to and during the first trial. Future studies could explore other paradigms that measure more "automatic" or "spontaneous" categorization, for example by making social categories seem irrelevant to the current task (Yang et al., n.d.). Some possible ways to do this include using a one-trial version of this task with more participants or hiding the gender/race questions among a variety of other questions (e.g., in intermixed trials, inserting filler questions like "How many faces were happy/sad?", "What is the average age of the faces?", or "Did you see a face at this location of the screen?").

Future studies could also use a more continuous display of total faces (e.g., from 8 to 20 total faces continuously; in our study each trial had 8, 10, 12, or 14 faces) and include more mixed displays (displays with both mixed-gender and mixed-race faces; in our study each trial involved a mixed-gender single-race display or mixed-race single-gender display). Moreover, providing cues about which subset to focus on (i.e., which gender or race will be asked) before the stimuli onset could tell us whether selective attention can be tuned to a subcategory in advance. These cues might be especially useful when more than two subsets of faces are involved (e.g., if all four subgroups, i.e., White male, White female, Black male, Black female were in the display). By comparing performances with and without cues, one can test how many subsets of faces are effectively encoded (i.e., using the point where performance did not differ between cued- and uncued-conditions). In addition, this task can also be used to measure categorization ability of other races (e.g., Asian) and other social categories (e.g., age).

It would be also interesting to explore whether the gender/race of the participants affected their performance on gender and race categorization in our studies. For example, would participants of a certain gender/race perform better? Would they perform better when the question is about their own gender/race or the other gender/race? In our studies, although we had a more racially diverse undergraduate sample (in Studies 1, 2, and 6) than online MTurk sample, the majority of the participants were White/European-American (n = 306, 66%), with each study including only a small number of individuals from other racial or ethnic groups. Given this, we were not able to perform analyses focusing on these questions. We were able to exploratorily examine gender-based subsetting, but did not find strong evidence for differences in this regard. Future studies could look at how categorization abilities might be affected by observer's gender and race, especially in relation to the other-race effect (ORE; performance is better for own-race than other-race faces on memory and perception, see Malpass & Kravitz, 1969; and see Thornton et al., 2019 for results that show that other-race faces are weighted more heavily than own-race faces).

Another fruitful avenue for future research is to explore the mechanisms of social categorization and its relationship with prejudice. Across six studies we showed that people are reasonably good at estimating the number of gendered and racial faces from briefly presented arrays of faces and that this ability is retained as stimuli are degraded in various ways. But what does this ability mean? Clearly, it helps us encode complex social information. But at the same time, it might also serve as a precursor to prejudice and stereotypes against certain groups (Dunham & Degner, 2013). This is because logically speaking, prejudice and stereotypes are only possible after the target groups are identified and categorized (e.g., due to links between categorization ability and prejudice; see Lee, Quinn, & Pascalis, 2017). Hence, it is possible that individuals who extract race or gender information more reliably might also be more likely to deploy automatic forms of prejudice. That is, are individual differences in sensitivity to gendered or racialized faces, or experiences with those social categories, related to the extent of prejudice, particularly automatically activated forms of prejudice? Future studies could build on our findings to look into these questions more closely.

In closing, across six studies we provided evidence that people can categorize and enumerate a crowd of gendered and racial faces and they can do so even in the face of significant perceptual degradations to the target faces. These abilities suggest that social categories and the cues that denote them reflect a powerful form of perceptual expertise that operates rapidly and even in the face of noisy input and distractor categories. In addition to being interesting as a perceptual process, these abilities have implications for person perception and prejudice formation because category-based perception is a necessary precursor to many forms of social bias.

Open practices

The studies in this article earned Open Materials, Open Data, and Preregistered badges for transparent practices. Materials are available at https://chicagofaces.org/default/download/ and names for the

specific stimuli we used at https://osf.io/5ka94/?view_only=15c365b391ed4e4d9d38c811ea359392. Data and analyses code are available at https://osf.io/5ka94/?view_only=15c365b391ed4e4d9d38c811ea359392. Preregistrations for each study are available at http://aspredicted.org/blind.php?x=nu24a8 (Study 1 and 2), http://aspredicted.org/blind.php?x=yi56xx (Study 1 and 2 replication), http://aspredicted.org/blind.php?x=mw9ek9 (Study 3), http://aspredicted.org/blind.php?x=eu7zw2 (Study 4), http://aspredicted.org/blind.php?x=6iz28u (Study 5), and http://aspredicted.org/blind.php?x=de4bw2 (Study 6).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2019.103893.

References

Alt, N. P., Goodale, B., Lick, D. J., & Johnson, K. L. (2017). Threat in the Company of Men: Ensemble perception and threat evaluations of groups varying in sex ratio. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550617731498 1948550617731498.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327.

Baudouin, J.-Y., & Humphreys, G. W. (2006). Configural information in gender categorisation. *Perception, 35*(4), 531–540. https://doi.org/10.1068/p3403.

Brown, E., & Perrett, D. I. (1993). What gives a face its gender? *Perception, 22*(7), 829–840. https://doi.org/10.1068/p220829.

Caharel, S., Montalan, B., Fromager, E., Bernard, C., Lalonde, R., & Mohamed, R. (2011). Other-race and inversion effects during the structural encoding stage of face processing in a race categorization task: An event-related brain potential study. *International Journal of Psychophysiology, 79*(2), 266–271. https://doi.org/10.1016/j.ijpsycho.2010.10.018.

Cordes, S., Goldstein, A., & Heller, E. (2014). Sets within sets: The influence of set membership on numerical estimates. *Journal of Experimental Psychology. Human Perception and Performance, 40*(1), 94–105. https://doi.org/10.1037/a0034131.

Dunham, Y., & Degner, J. (2013). From categories to exemplars (and back again). In M. R. Banaji, & S. A. Gelman (Eds.). *Navigating the social world: What infants, children, and other species can teach us* (pp. 275–280). .

Dupuis-Roy, N., Fortin, I., Fiset, D., & Gosselin, F. (2009). Uncovering gender discrimination cues in a realistic setting. *Journal of Vision, 9*(2), 10. https://doi.org/10.1167/9.2.10.

Freeman, J. B., Pauker, K., Apfelbaum, E. P., & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology, 46*(1), 179–185. https://doi.org/10.1016/j.jesp.2009.10.002.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception, 29*(2), 159–170. https://doi.org/10.1068/p3012.

Goodale, B. M., Alt, N. P., Lick, D. J., & Johnson, K. L. (2018). Groups at a glance: Perceivers infer social belonging in a group based on perceptual summaries of sex ratio. *Journal of Experimental Psychology: General, 147*(11), 1660–1676. https://doi.org/10.1037/xge0000450.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 718–734. https://doi.org/10.1037/a0013899.

Haberman, J., & Whitney, D. (2011). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe, & L. Robertson (Eds.). *From perception to consciousness: Searching with Anne Treisman*. Oxford University Press.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science, 17*(7), 572–576. https://doi.org/10.1111/j.1467-9280.2006.01746.x.

Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: The role of shape and colour. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 261*(1362), 367–373. https://doi.org/10.1098/rspb.1995.0161.

Horne, E. P., & Turnbull, C. E. (1977). Variables of color, duration, frequency, presentation order, and sex in the estimation of dot frequency. *The Journal of General Psychology, 96*(1), 135–142. https://doi.org/10.1080/00221309.1977.9920807.

Jung, W., Bülthoff, I., & Armann, R. G. M. (2017). The contribution of foveal and peripheral visual information to ensemble representation of face race. *Journal of Vision, 17*(13), 11. https://doi.org/10.1167/17.13.11.

Lamer, S. A., Sweeny, T. D., Dyer, M. L., & Weisbuch, M. (2018). Rapid visual perception of interracial crowds: Racial category learning from emotional segregation. *Journal of Experimental Psychology: General, 147*(5), 683–701. https://doi.org/10.1037/xge0000443.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 53*(2), 513–536. https://doi.org/10.1080/713755889.

Lee, K., Quinn, P. C., & Pascalis, O. (2017). Face race processing and racial bias in early development: A perceptual-social linkage. *Current Directions in Psychological Science, 26*(3), 256–262. https://doi.org/10.1177/0963721417690276.

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

Malpass, R., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology, 13*(4), 330–334.

Martin, D., Swainson, R., Slessor, G., Hutchison, J., Marosi, D., & Cunningham, S. J. (2015). The simultaneous extraction of multiple social categories from unfamiliar faces. *Journal of Experimental Social Psychology, 60*, 51–58. https://doi.org/10.1016/j.jesp.2015.03.009.

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Nestor, A., & Tarr, M. J. (2008). Gender recognition of human faces using color. *Psychological Science, 19*(12), 1242–1246. https://doi.org/10.1111/j.1467-9280.2008.02232.x.

Phillips, L. T., Slepian, M. L., & Hughes, B. L. (2018). Perceiving groups: The people perception of diversity and hierarchy. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000120.

Stevenage, S. V., & Osborne, C. D. (2006). Making heads turn: The effect of familiarity and stimulus rotation on a gender-classification task. *Perception, 35*(11), 1485–1494. https://doi.org/10.1068/p5409.

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in Cognitive Sciences, 5*(5), 211–215. https://doi.org/10.1016/S1364-6613(00)01626-0.

Tarr, M. J., Kersten, D., Cheng, Y., & Rossion, B. (2001). It's pat! Sexing faces using only red and green. *Journal of Vision, 1*(3), 337. https://doi.org/10.1167/1.3.337.

Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research, 51*(11), 1273–1278. https://doi.org/10.1016/j.visres.2011.04.002.

Thornton, I. M., Srismith, D., Oxner, M., & Hayward, W. G. (2019). Other-race faces are given more weight than own-race faces when assessing the composition of crowds. *Vision Research, 157*, 159–168. https://doi.org/10.1016/j.visres.2018.02.008.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face. *British Journal of Psychology, 79*(4), 471.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x.

Yang, X., Langfus, J., Halberda, J., & Dunham, Y. (2019). *Spontaneous encoding of gender and race in visual working memory: Evidence from a change detection paradigm.* [under review] under review.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141–145.

Yip, A. W., & Sinha, P. (2002). Contribution of color to face recognition. *Perception, 31*(8), 995–1003. https://doi.org/10.1068/p3376.

Zarate, M. A., & Smith, E. R. (1990). Person categorization and stereotyping. *Social Cognition, 8*(2), 161–185. https://doi.org/10.1521/soco.1990.8.2.161.