Behav Res (2011) 43:224–228
DOI 10.3758/s13428-010-0029-9

A computer-generated face database with ratings on realism, masculinity, race, and stereotypy

Heath E. Matheson · Patricia A. McMullen

Published online: 7 December 2010

Abstract Ratings of realism, masculinity, race, and racial stereotypy were collected on a set of computer-generated faces representing European, South East Asian, and African American ethnicities. To determine if these faces are processed in the same way as photographs of real faces, we demonstrated with these faces superior memory performance for upright faces over inverted faces (the face inversion effect). Further, in observers of European decent, we found both superior memory for European faces and a larger inversion effect for European than African American faces. Based on these results, we believe that this set of faces may be of use in perceptual investigations in which race is a critical manipulation.

Keywords Face perception · Other-race effect · Face database

Introduction

Behavioral and neuroscientific investigations of face perception often use databases of static face images to explore the processing of facial identity, emotion recognition/classification, gender classification, and the effects of face race on both perceptual and socio-cognitive variables of interest (see Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2000 for reviews of behavioral and neuroscientific investigations of identity, emotion and gender processing; see Meissner & Bringham, 2001 for behavioral review of race processing). Some databases available to researchers include: the CAS-PEAL, the FERET, the MIT, the YALE, and the Korean Face Database and the Japanese Female Facial Expression Database (see Gross, 2005, details of a wide selection of faces). Critically, these databases are often poorly controlled with respect to image quality, distinctive facial features, and external facial paraphernalia (e.g., hair). Further, to date there is no freely available database which systematically represents faces of multiple races. Historically, researchers are often interested in the difference between races of high familiarity (often faces of one's own race) and races of low familiarity (usually a single race that is different from the participant). However, this traditional comparison confounds participant experience (high vs. low experience) and low-level structural features (contrast, skin tone differences etc). To obviate this, the use of three races is necessary. Specifically, we sought to develop a face data base that contains large and equal numbers of three different races. With three races, an observer may be highly familiar with one and unfamiliar with the others. In this circumstance, three races allows for the effects of race-specific structural features to be compared in the absence of experiential differences (other-race A vs. other-race B), and for determination of the combined effects of experience and any structural differences (own-race vs. other-race); further, in some cases observers may be highly familiar with two of the represented races but not the third. In this case, experience can be held constant while comparing the effects of structural differences. In experiment 1, we used 3-D computer-generation software to generate faces of three different races. This program randomly combines facial features to create schematic faces that are close representations of real faces. Because the faces are schematic representations and so ecological validity is a concern, we

Electronic supplementary material The online version of this article (doi:10.3758/s13428-010-0029-9) contains supplementary material, which is available to authorized users.

H. E. Matheson (✉) · P. A. McMullen
Department of Psychology/Neuroscience, Dalhousie University, Halifax, Nova Scotia B3H 4J1, Canada
e-mail: heathmatheson@dal.ca

have collected ratings on a number of critical variables on each face, including facial masculinity, facial realism, stereotypy, and a categorical response regarding face-race. Using these ratings, researchers will be able to select subsets of the faces that are matched on one (or all) of these critical variables. To further address concerns about ecological validity, we have replicated robust phenomena in the face literature using the computer generated faces in a second experiment. First, we show superior memory for upright compared to inverted faces (i.e., the face inversion effect; Yin, 1969; see Rossion & Gauthier, 2002, for a review), second we show better recognition memory for own-race than other-race faces (i.e., the cross-race effect; Lindsay, Jack, & Christian, 1991), and finally we show a larger effect of inversion for own-race faces than other-race faces (the interaction between the two; Rhodes, Brake, Taylor, & Tan, 1989).

Material and methods

Experiment 1

In Experiment 1 we obtained observer ratings on a number of important variables.

Participants

Seventy-four participants (16 males, mean age = 21.5, *SD* = 3.7) were recruited from the psychology department at Dalhousie University. All gave written informed consent to participate and all had normal or corrected-to-normal vision.

Stimuli

Eighty face-images of three different races were generated using FaceGen Software (Singular Inversions, VA) (see Fig. 1 for examples). Images were rendered in greyscale. Facial images subtended approximately 10.47 degrees of visual angle and appeared on a black background. In two pilot sessions, the experimenter and a research assistant replaced any face that was too unrealistic or too feminine based on mutual agreement. Each face image was then paired with one of three questions and a seven-point Likert scale (appearing below the image): (A) How realistic is this face? (B) How masculine is this face? (C) How good of an exemplar of the race is this face? (A rating of one was considered as 'low' and seven as 'high'.) A fourth question, (D) "What race is this face?" was paired with the choice of four answers: (A) European, (B) South East Asian, (C) African American, (D) None of the above.

Procedure

To ensure that each face received approximately the same number of ratings (because of the constraint of time), each participant was randomly assigned to one of four subsets of images (each image was rated by 36–38 participants). Questions were displayed in blocks. Further, within each subgroup, subjects were pseudorandomly assigned to one of six orders of questions.

On each trial, a face image appeared on the screen above the question and Likert scale and remained on the screen until participants made a response with the number pad. Participants were instructed to respond quickly but thoughtfully and were encouraged to use the full width of the seven-point scale.

Results and discussion

Two participants were excluded because they had participated in other face recognition experiments in our lab. Three participants failed to stay within the seven-point Likert-scale range on at least one of the trials. These occurred on < .01 % of the trials and were excluded from the final analysis. The mean response and standard error of ratings for each question for each race is presented in Table 1.

A 3 (race) X 3 (question) analysis of variance (ANOVA) was conducted to explore whether there were any system-

Fig. 1 Examples of the colored face stimuli generated by FaceGen

atic differences across the ratings of the three races within each question (the categorization data was not included in this analysis). The analysis revealed a main effect of race $F(2, 142) = 44.44$, $MSE = .578$, $p < 0.001$, a main effect of question, $F(2, 142) = 11.59$, $MSE = .77$, $p < 0.001$, and an interaction, $F(4, 284) = 16.4$, $MSE = .163$, $p < 0.001$.

To look at these differences more closely, post hoc t-tests were conducted. For ratings of masculinity, African American faces were not rated as more masculine ($M = 5.15$) than European faces ($M = 5$), $p = 0.145$, though both were rated more masculine than South East Asian faces ($M = 4.35$), $ps < 0.01$. This demonstrates that African American and European faces were rated consistently higher for masculinity. For ratings of realism, a slightly different pattern was seen. African American faces were consistently rated as more realistic ($M = 4.92$) than both the European faces ($M = 4.36$) and the South East Asian faces ($M = 4.36$), $ps < 0.001$, with no difference between the latter two, $p = 0.9$. Here the African American faces are perceived as more real than the other races. The ratings for stereotypy follow the same pattern. Specifically, African American faces were rated as more stereotypical ($M = 5.38$) than both the European faces ($M = 4.74$) and the South East Asian faces ($M = 4.70$), $ps < 0.001$, though the latter two were not different, $p = 0.68$.

Overall, these patterns suggest that the computer-generated African American images are more stereotypical and masculine. Importantly, the ratings will allow researchers to equate subsets of stimuli to eliminate these biases. To the best of our knowledge, this is the only computer-generated face set that includes these types of ratings.

For the question "Which race?", a '1' corresponded to identifying the race as European, '2' as South East Asian, and '3' as African American. Race had a significant effect on this question, $F(2, 142) = 1709.46$, $p < 0.001$. Post hoc tests revealed differences between all comparisons, $ps < 0.001$. Importantly, there is a small amount of variability around the means (European, $M = 1.25$, $SE = .036$; South East Asian, $M = 2.11$, $SE = .019$; African American, $M = 3.023$, $SE = .09$), demonstrating that categorization was not perfect. This is important as it suggests that not all participants categorized all the faces similarly, though some of these ratings may have been accidental.

Together, the database shows some systematic difference between ratings across all 80 faces within each race. For the first time, researchers will be able to use each of these ratings to develop subsets that are matched on any (or all) of the variables of interest.

Experiment 2

Though participants were able to categorize the faces as belonging to one of three races, there is no *direct* evidence from the rating experiment that the visual system processes these artificial face stimuli in the same way as real face stimuli. Indeed participants could have based their judgments on low level stimulus features such as contrast or perhaps single features within each face. To determine whether the visual system responds to these computer-generated faces in a manner similar to real faces we attempted to replicate the well known interaction between face inversion and face race in a short memory experiment (Rhodes, Brake, Taylor, & Tan, 1989).

Participants

Seventeen Caucasian participants (eight males, age range = 17–24) were recruited from the psychology department at Dalhousie University. All gave written informed consent to participate and all had normal or corrected-to-normal vision.

Stimuli

Forty faces from both the European and African American sets were randomly selected. Thus, there were ten study and ten distractor faces in the upright condition and ten study and ten distractor faces in the inverted condition.

Procedure

The testing procedure was similar to that of the Recognition Memory Test used in neuropsychological evaluations (Warrington, 1984). Participants were informed that they were completing a memory test for faces. On each trial, a study face appeared for 2 s. Participants were instructed to decide whether they thought the face was pleasant, by verbal report. An experimenter stood in the room to ensure that participants verbally reported their decisions. These were not recorded. Immediately following the study phase, a test phase began in which study and distractor faces

Table 1 Mean ratings and the standard error for the four ratings

	Masculinity		Realism		Stereotypy		Which Race	
Race	*Mean*	*SE*	*Mean*	*SE*	*Mean*	*SE*	*Mean*	*SE*
African American	5.15	.087	4.92	.126	5.38	.122	3.02	.01
European	5.01	.087	4.36	.129	4.74	.114	1.25	.04
South East Asian	4.35	.087	4.36	.117	4.70	.117	2.11	.02

(randomly presented) were displayed, and participants were to press 'z' for previously seen faces, or 'm' for novel faces (counterbalanced). The study faces remained on the screen until participants made a response. Each participant completed tests with both upright and inverted European and African faces.

Results and discussion

A 2 (race) X 2 (orientation) ANOVA on *d'* (sensitivity) measures revealed a main effect of race, $F(1, 16) = 6.89$, $MSE = .45$, $p = 0.018$, demonstrating better recognition of European faces ($M = .79$, $SE = .126$) than African American faces ($M = .363$, $SE = .109$) overall. There was also a significant effect of orientation, $F(1, 16) = 17.23$, $MSE = .743$, $p < 0.01$, demonstrating better recognition of upright faces ($M = 1.01$, $SE = .15$) than inverted faces ($M = .142$, $SE = .12$) overall. Importantly, the analysis revealed a significant interaction, $F(1, 16) = 7.44$, $MSE = .436$, $p = 0.015$. Post hoc *t*tests revealed that this effect was due to a difference between upright European ($M = 1.44$, $SE = .26$) and upright African American ($M = .58$, $SE = .11$) faces, but no significant difference between inverted European ($M = .14$, $SE = .14$) and inverted African American ($M = .15$, $SE = .16$) faces (see Fig. 2).

These results replicate both the inversion effect, superior memory for own-race faces, and have shown the characteristic interaction between the two. These results strongly suggest that the visual system treats the computer-generated own-race and other-race faces used in this experiment in a similar way to real grey-scale, photographic face images. This encourages the use of the face data base in further experimental testing. However, two important points need to be made. First, recognition memory performance is low overall, suggesting that recognition is difficult even when the study set only contains ten faces (this is especially seen in both inverted conditions, which might reflect a floor effect). Further, almost all participants reported finding this task challenging. We speculate that the face set poses challenges for the face processing system because all of the faces are visually similar; that is, they are all similar in shape and arrangement of features, and there are no distinctive elements to each image (i.e all faces are bald, without skin imperfections etc.). Despite this, we encourage the use of the database in further experiments, especially those that require challenging perceptual discriminations or memory tests.

General discussion

We have successfully developed a large computer-generated face database that represents three different races. Further, we have collected ratings on a number of critical variables that should be of interest to researchers using these faces and will allow researchers to match subsets of faces on these variables. Further, we have shown three behavioral phenomena that have been reported in the perception/memory literature on face-race processing, namely the inversion effect, the other-race effect, and an interaction between the two (e.g., Rhodes et al., 1989). Importantly, the face images are all of the same quality, they do not possess distinctive features, and external paraphernalia is absent. Further, because three races are represented in the database researchers can compare the effects of experi-

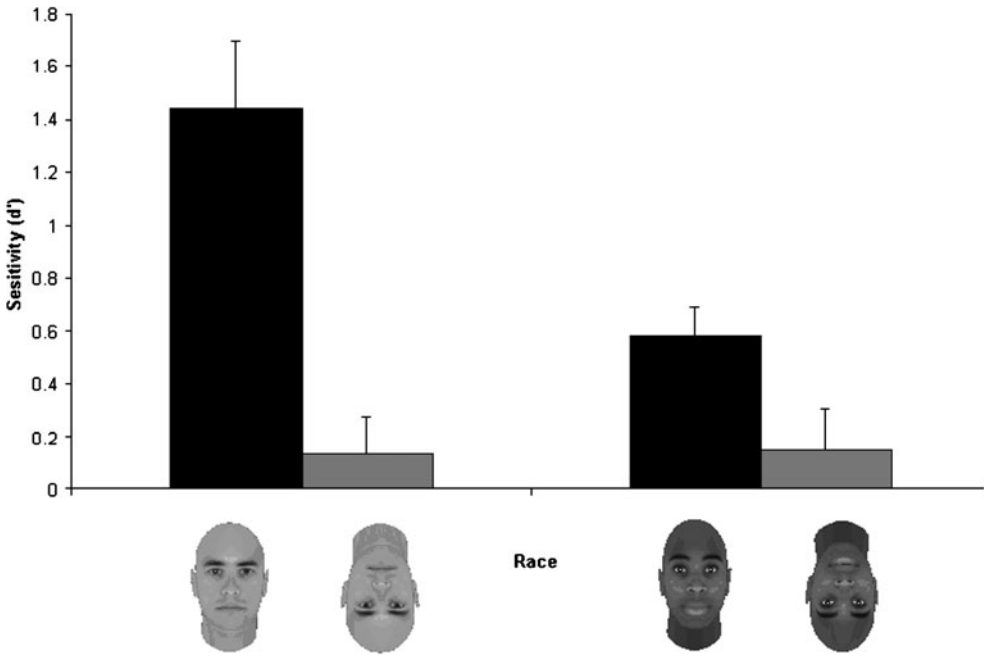

Fig. 2 Mean d' as a function of condition. From *left to right*: upright European, inverted European, upright African American, and inverted African American. *Error bars* represent standard error of the mean

ence (own-race vs. other-race faces) and/or the effects of systematic differences between faces of two unfamiliar races (other-race face A and other-race face B). Although the use of highly similar, bald, computer-generated faces might be expected to reduce the ecological validity of findings based on the use of these stimuli, our experimental work has shown otherwise and we anticipate that many researchers will find this database useful for a wide range of perceptual and recognition experiments.

Acknowledgements We would like to thank Kayla Hyland for her help in organizing rating data. Funding provided by NSERC PGS D to Heath Matheson. The authors report no conflicting interests.

References

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

Gross, R. (2005). Face databases. In S. Li & A. Jain (Eds.), *Handbook of Face Recognition*. Springer.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223–233.

Lindsay, D. S., Jack, P. C., & Christian, M. A. (1991). Other-race face perception. *The Journal of Applied Psychology, 76*(4), 587–589.

Meissner, C. A., & Bringham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology Public Policy and Law, 7*(1), 3–35.

Rhodes, G., Brake, S., Taylor, K., & Tan, S. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology, 80*, 313–331.

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioural and Cognitive Neuroscience Reviews, 1*(1), 63–75.

Warrington, E. K. (1984). *Recognition memory test*. Berkshire: NFER-Nelson.

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