Behav Res (2015) 47:1122–1135
DOI 10.3758/s13428-014-0532-5

The Chicago face database: A free stimulus set of faces and norming data

Debbie S. Ma · Joshua Correll · Bernd Wittenbrink

Published online: 13 January 2015

Abstract Researchers studying a range of psychological phenomena (e.g., theory of mind, emotion, stereotyping and prejudice, interpersonal attraction, etc.) sometimes employ photographs of people as stimuli. In this paper, we introduce the Chicago Face Database, a free resource consisting of 158 high-resolution, standardized photographs of Black and White males and females between the ages of 18 and 40 years and extensive data about these targets. In Study 1, we report pretesting of these faces, which includes both subjective norming data and objective physical measurements of the images included in the database. In Study 2 we surveyed psychology researchers to assess the suitability of these targets for research purposes and explored factors that were associated with researchers' judgments of suitability. Instructions are outlined for those interested in obtaining access to the stimulus set and accompanying ratings and measures.

Keywords Face database · Multiracial faces · Normed face stimuli

Electronic supplementary material The online version of this article (doi:10.3758/s13428-014-0532-5) contains supplementary material, which is available to authorized users.

D. S. Ma (✉)
Department of Psychology, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330, USA
e-mail: debbie.ma@csun.edu

J. Correll
University of Colorado at Boulder, Boulder, CO, USA

B. Wittenbrink
University of Chicago, Chicago, IL, USA

Introduction

Faces occupy privileged status in human psychology. Minutes after birth human neonates preferentially orient toward schematic faces (Morton & Johnson, 1991; Pascalis & Kelly, 2009) and by adulthood humans show a remarkable aptitude for memorizing, attending to, and recognizing faces (Bruce & Young, 1986; Coin & Tiberghien, 1997; Sato & Yoshikawa, 2013; Theeuwes & Van der Stigchel, 2006). Adults also spontaneously ascribe traits to faces within a matter of milliseconds (Willis & Todorov, 2006) and research suggests that these rapid judgments correlate with actual personality (e.g., Penton-Voak, Pound, Little, & Perrett, 2006). The social and biological significance of faces is so great that humans have an area of visual cortex that appears to have evolved to subserve face processing (Kanwisher, McDermott, & Chun, 1997; Sergent, Ohta, & MacDonald, 1992).

Given the significance of faces to basic human processes, many of the research paradigms used to study theory of mind, impression formation, spontaneous trait inference, group processes, interpersonal attraction, aggression, stereotyping and prejudice, emotions, etc., involve the presentation of face stimuli to participants. The use of faces can be incidental, such as when an experimenter needs to convince participants that they are in a group of other participants (e.g., Ratner, Kaul, & Van Bavel, 2012; Williams & Jarvis, 2006), or pivotal to the research question. For example, Baron-Cohen and colleagues' (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) "Mind in the Eyes" test in which researchers present participants with a desaturated image cropped to only show the eye area of a face and participants are asked to indicate which mental trait best describes the apparent expression captured in the eyes. Other examples involve paradigms used to implicitly measure attitudes toward racial groups in which participants are presented with face stimuli from different racial groups

and asked to categorize or respond to these stimuli in some way (e.g., Fazio, Jackson, & Dunton, 1995; Greenwald & Banaji, 1995).

Stimuli collection for these experimental procedures varies widely. Researchers may elect to take their own photographs of targets and process those images in-house (e.g., Correll, Park, Judd, & Wittenbrink, 2002), identify usable targets from archives, such as yearbooks (e.g., Blair, Judd, & Fallman, 2004), find images from online or print sources (e.g., Baron-Cohen et al., 1997), create computer-generated faces (e.g., Todorov, Pakrashi, & Oosterhof, 2009), or use stimuli from published databases, such as NimStim or Project Implicit (e.g., Ma & Devos, 2013; McConnell & Leibold, 2001). The amount of effort required to gather, standardize, and pre-test pictorial stimuli can be daunting. Specifying selection criteria, finding people to be photographed, taking the photographs, and processing and pretesting the images might take months or years. For this reason, databases of faces provide a convenient and attractive alternative to these hassles. Beyond the practical reasons, databases offer a number of benefits. Common databases may facilitate comparisons across studies, allow for and promote exact replications, and improve experimental control. These advantages seem all the more valuable in light of recent calls for replication and better methodological practices within the field (Asendorpf et al., 2013; Kahneman, 2012; Yong, 2012). In this paper, we introduce a new database of face stimuli, which aims to address several shortcomings of existing public-use stimulus sets.

Existing databases

Several databases of face stimuli exist. Though not exhaustive, we describe a few commonly used databases, which represent distinct techniques that people have employed when developing stimulus sets. One type of database reported in the literature involves photographs of targets making a variety of facial expressions. The Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) is one such example. It is comprised of 70 White amateur actors (35 male, 35 female), ranging from 20 to 30 years old, and wearing a standard gray shirt. Each individual was photographed making different emotional expressions, and photographs were also taken at different angles. The database contains 4,900 unique images in total, carefully standardized and controlled. Researchers can use the KDEF for research purposes at no cost after agreeing to the terms of use. Like KDEF, some other databases use actors (e.g., Tottenham et al., 2009), but the majority employ lay volunteers (e.g., Belhumeur, Hespanha, & Kriegman, 1997; Minear & Park, 2004; Rhodes, Proffitt, Grady, & Sumich, 1998; Thomaz & Giraldi, 2010). A variant of this type of database includes the Max Planck Institute Head Database (MPI Head Database; Troje & Bülthoff, 1996). The MPI Head Databases employed laser capture technology to record 3D information about the shape and coloring of 200 male and female White Europeans from seven different angles. From these scans, researchers generated five sets of morphed targets that can be used for research purposes. This morphing was done to preserve the identity of the individuals who were scanned.

A different category of face databases has been assembled by researchers who are interested in developing facial recognition technology (e.g., Berg et al., 2004; Kumar, Belhumeur, & Nayar, 2008; Rowley, Baluja, & Kanade, 1998). Berg and colleagues gathered 30,281 images from various media outlets in order to test the accuracy of their program. As a result, their database – Faces in the Wild – contains many well known public figures. These images vary along a number of dimensions, including saturation, size, resolution, lighting conditions, facial expressions, clothing, eye gaze, and more. As with many published databases, researchers must agree to certain terms in order to utilize Faces in the Wild before accessing this free resource.

A third type of databases is comprised of artificial faces (e.g., Matheson & McMullen, 2011). These faces can be computer generated using software programs, morphed from real faces, or sketched. One such resource containing morphed faces is Project Implicit (www.implicit.harvard.edu). Targets include men, women, Blacks, Whites, Native Americans, kids, elderly persons, Asians, and more. For the most part these stimulus sets contain a small number of targets (between five and ten targets per category).

Limitations of existing databases

Lack of information about targets Many, though not all, of the images in published face databases have been validated in some way or another. Faces in NimStim, for instance, have been validated for their emotional expressions by a sample of undergraduate participants. Ekman and Friesen's (1976) Pictures of Facial Affect (POFA) consists of 110 different emotional expressions, which have been validated in a variety of samples. By contrast, the validation of stimuli gathered using face detection software tends to be carried out by a single individual whose responsibility is to verify that the software accurately returns faces and not non-faces. Generally speaking, researchers have considered a relatively narrow set of dimensions when validating their targets. This may be because the databases are byproducts of specific research questions (e.g., the researcher may be studying emotion and only collects data on the emotional quality of their targets). Ultimately this means that little else is known about how the targets are perceived on other psychologically meaningful (and possibly consequential) dimensions. A broader understanding of the faces' underlying stimulus attributes could provide researchers with more guidance when selecting stimuli and interpreting results. The lack of comprehensive

norming data about face stimuli is one of the primary motivations for the current undertaking.

A useful model for normed stimuli comes from the International Affective Picture System (IAPS; Lang, Bradley, & Curthbert, 1997), a large database of standardized color images. IAPS includes 700 images as well as information about how each image is perceived with respect to affective valence, arousal, and dominance/control. Ratings data were obtained from 100 college students as well as from smaller samples of children (7–14 years old). As we described above, these data can improve experimental control and may even enable and motivate novel research pursuits.

Demographic homogeneity of targets In addition, the currently available face stimuli have several other limitations, which may have important practical and theoretical implications. One limitation involves a high level of homogeneity. The majority of the databases are comprised of individuals of European descent (e.g., Ekman & Friesen, 1976; Lundqvist et al., 1998; Troje & Bülthoff, 1996). Other databases such as NimStim (Tottenham et al., 2009) may include Asian, Black, White, and Latino targets; however, the total number of unique individuals within each of these categories is quite low. In addition, individuals featured in NimStim and the KDEF are all in their twenties. Demographic restrictions of this sort become even more problematic if a researcher hopes to control for additional target characteristics (e.g., equate White and Black targets on attractiveness), cross target factors (e.g., find equal numbers of male and female White and Black targets), or constrain their search criteria (e.g., isolate Black females). What initially seems like a reasonably sized stimulus set can quickly diminish to a small, restrictive set of options. If researchers combine faces from multiple databases (e.g., Ackerman et al., 2006) the stimuli often vary in terms of quality (lighting, clothing, resolution, etc.).

Stimulus homogeneity raises critical theoretical concerns. In particular, research based on stimuli with artificially low variation may overestimate or underestimate effects and miss important moderators (Fiedler, 2011). For example, research suggests that feature-based variation within social categories can predict attitudes and stereotyping over and above category membership (Blair et al., 2004; Livingston & Brewer, 2002; Ma & Correll, 2011). Research on feature-based prejudice finds that Black stimuli with more Black prototypic features are judged more negatively whereas White stimuli with more White prototypic features are judged more positively. In the absence of variability among the stimuli on the dimension of prototypicality, it would be impossible to detect feature-based prejudice. Conceptually similar issues may arise with respect to dimensions other than race. Researchers investigating affect, for example, have demonstrated that variation in the intensity of an emotional expression produces corresponding activation differences in neural responsivity in areas of the brain related to emotion processing (Morris et al., 1998). As with race, failing to account for variability in emotional intensity – or really any stimulus characteristic – could lead the researcher to unwittingly draw spurious conclusions. A second and related theoretical concern is that researchers may overlook meaningful boundary conditions. For example, if faces in a database include only younger adults, researchers may fail to detect effects that would be observed in response to children or older adults (Steffensmeier, Ulmer, & Kramer, 1998). A third issue relates to ecological validity. As a general rule, conclusions based on heterogeneous stimuli tend to be less idiosyncratic (i.e., findings should be less stimuli-dependent given greater diversity) and thus should be more likely to generalize to other stimuli. Stimulus sampling poses real threats to construct validity, potentially affecting both external and internal validity (Judd, Westfall, & Kenny, 2012; Wells & Windschitl, 1999).

Stimulus standardization and quality Although heterogeneity on some dimensions may be valuable as described above, researchers may need for stimuli to be more homogeneous on other dimensions. Databases vary widely in terms of image quality and consistency. Stimuli gathered from online or media sources tend to be extremely variable in terms of dress, lighting, target gaze, image resolution, etc., whereas databases of photographs taken in a studio tend to be fairly consistent in terms of photo quality. However, even for studio-based databases, there is a great deal of variability *between* databases.

Goals of the current research

The current paper describes the development of a new database of facial stimuli: the Chicago Face Database (CFD). Critically, this database involves several unique features intended to fill a specific niche for experimental researchers. Our goal was to create a free resource for the scientific community that addresses as many of the issues raised above as possible. To respond to the issue of demographic homogeneity, we included a large number of carefully pretested White and Black male and female faces. Here, we report the first phase of stimuli and data collection, covering a subset of 158 Black and White male and female faces included in the database. To address the issue of standardization, the stimuli are carefully controlled and produced in high-resolution image files. Perhaps the most important – and most unique – aspect of the CFD is that we address the issue of information about the stimuli. Extensive norming data accompany each individual target included in the database. Like IAPS, we provide users with a variety of information about each target in the database. As reported in Study 1, each target was rated by a large sample of participants on a number of psychologically meaningful dimensions (e.g., babyfacedness and attractiveness). We also took physical measurements of the faces,

providing objective data. In Study 2 we solicited input from experts within the field of psychology to rate the suitability of the CFD targets for research purposes. All images along with averages and standard deviations of the ratings obtained in Studies 1 and 2 are available for download at no cost at http://www.chicagofaces.org/.

Study 1: Establishing the Chicago Face Database

The first phase of developing the CFD involved collecting stimuli and gathering data about each target. We took high-resolution, digital photographs of targets displaying a variety of facial expressions under standardized conditions (e.g., lighting, face angle, eye level, etc.). Next, we submitted photographs of the targets to extensive testing. In order to obtain norming data, we collected subjective ratings of the targets from a large sample of participants. We also obtained objective physical measures of the faces and submitted these to factor analysis in order to examine covariation among facial features.

Method

Collecting stimuli and stimuli standardization

Target sample Individuals were recruited from the Chicago Research Laboratory of the University of Chicago Booth School of Business, located in downtown Chicago. Potential volunteers were contacted via email to serve as targets for the development of a database of faces to be used for research purposes. During the recruitment process we also targeted amateur actors and used snowball sampling in order to obtain a pool of individuals whom we believed would be able to produce reliable and believable facial expressions. Volunteers were required to be between the ages of 18 and 40 years and to self-identify as Black or White. 177 targets agreed to be photographed for the database, of which 19 were excluded due to severe facial deformities or because the photo was not taken properly. Of the remaining 158 volunteers, 37 were Black males, 48 were Black females, 36 were White males, and 37 were White females. Upon arrival, participants were asked to carefully read a consent/release form, allowing us to use their photos for research purposes. Afterward, participants changed into a heather grey t-shirt (or wore it over their normal clothing) and were seated at a fixed distance from a digital camera. The camera height was adjusted to the target's eye level. Participants sat in front of a white cloth backdrop. To control for lighting conditions, three photo lamps were trained on the target. Two lights were trained on the front of the target, and the third (a hair lamp) was used to reduce shadowing. Participants were asked to make neutral, happy, threatening, and fearful expressions while maintaining a constant head position. At least three rounds of photographs were taken for each person making each of the facial expressions. In the first round, participants simply received specific verbal prompts (e.g., "Make a closed mouth smile"). In the second and third round of photos, the photographer again gave verbal prompts (e.g., "Make a closed mouth smile") for each of the desired facial expressions, but the prompts were given at a faster pace to elicit a more spontaneous expression. At times, the photographer followed up the general prompt with more specific directions (e.g., "Make sure to engage your eyes in the smile"). In cases where participants were unable to produce believable looking facial expressions, the photographer presented the individual with images of validated emotional expressions to give examples and provided them with a mirror (Ekman & Friesen, 1976). This resulted in multiple photographs for each target making each of the different facial expressions. Photographs were taken to include the shoulders and head using a 50-mm 1/8 f lens. The photographs were shot in high-resolution, raw format. Sessions lasted approximately 20 minutes and participants were compensated $20.

Stimuli standardization We selected one image of each target displaying a neutral facial expression based on how apparently neutral the face seemed and the head positioning of the person (i.e., we selected images in which the target's head was as vertical as possible and the face was turned directly toward the camera such that we could see both ears equally). Selection of the emotional expression images (fearful/afraid, angry, happy, closed mouth smile, open mouth smile) focused primarily on the quality and believability of the emotional expression and then on head positioning. To make these selections, two independent judges rated each of the emotional expression images in terms of how believable the expression was on a 1–9 Likert scale (1 = not at all believable; 9 = very believable). One image was selected for each of the emotional expressions per target. Neutral and emotional expression image files were edited using Adobe Photoshop. Digital modifications were performed to remove earrings, facial piercings, and facial hair. We also resized images such that the size of the core facial features as depicted in the photo was roughly equivalent across targets. The original dimensions of the photos were 3008 pixels (wide) ×2000 pixels (high). To standardize the size with which the faces are depicted in the photo an invisible 796 pixels (wide) ×435 pixels (high) rectangle was fit over the targets' core facial features such that the rectangle met one or both of the following conditions: (1) the vertical distance between the lowest part of the inner brow and the top of the upper lip corresponded to the height of the rectangle, or (2) the horizontal distance between the farthest visible extent of the cheek bones matched the width of the rectangle. The resulting photos

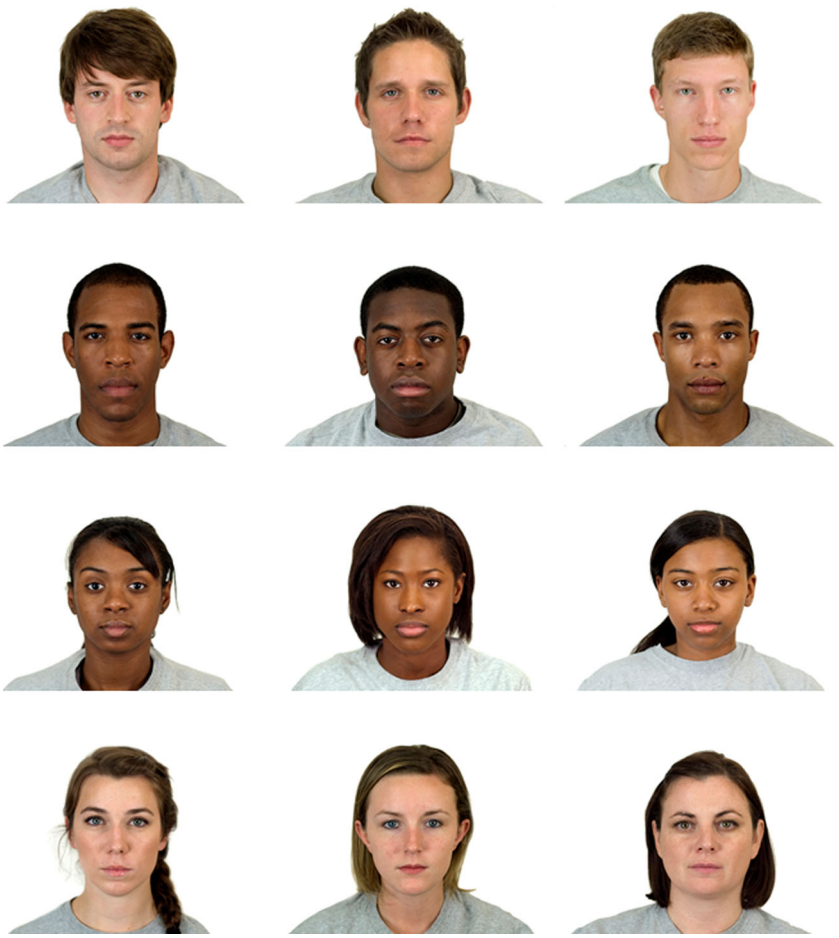

Fig. 1 Sample stimuli from the Chicago Face Database

were cropped to 2444 pixels (wide) ×1718 pixels (high). Finally, images were equated for color temperature and placed onto a plain white background (see Fig. 1 for sample stimuli).

Gathering norming data

Norming data were collected for the neutral-expression pictures. Standardized image files of each target making each of the emotional expressions are also available in the database, but have not been normed.

Subjective ratings A convenience sample of 1,087 raters made subjective ratings of the standardized neutral faces. Participants included 552 females and 308 males (227 did not report) and came from diverse racial backgrounds (516 White, 117 Asian, 74 Black, 72 biracial or multiracial, 57 Latino, 18 other, and 233 did not report). The average age of the sample was 26.75 (*SD* = 10.54). Raters were presented with the neutral image from each target one at a time on the computer using Qualtrics Research Suite Software. Participants were first asked to estimate the age of each target, categorize each target as either Asian, Black, Hispanic/Latino, White, or Other, and indicate the gender of each target. Next, participants rated each target in terms of how threatening, masculine, feminine, baby-faced, attractive, trustworthy, happy, angry, sad, disgusted, surprised, fearful/afraid, and unusual (would stand out in a crowd) they were. For this latter set of ratings, participants were instructed to consider each target in relation to others of the same race and gender when making each judgment. We opted to present raters with images from multiple target categories in order to maximize the comparability of the ratings across categories. Previous research has shown that people hold well formed mental representations of category prototypes for basic social categories such as race and gender (e.g., Blair & Judd, 2010; Zebrowitz, 1997) and should be able to complete these judgments across multiple categories. Participants responded on a 1–7 Likert scale (1 = Not at all, 7 = Extremely). Because it was not feasible to have raters judge each of the 158 targets on all 14 dimensions, participants rated 15 targets that were randomly selected for each rater to judge. The number of targets was reduced to ten after collecting 168 cases when we realized that participants required more time than anticipated to rate 15 faces on all 14 dimensions. We limited the number of targets rated, because we were concerned with participant fatigue and wanted to

ensure that raters were fully engaged while providing their judgments. Sampling was done without replacement and refreshed only after all of the 158 faces were judged. Eight $25 cash prizes were randomly awarded to raters who completed the survey. We also administered a second survey to another group of participants to assess participants' ratings of the targets in terms of how Eurocentric-Afrocentric each was on a 100-point Likert scale (0 = Very White/Eurocentric; 100 = Very Black/Afrocentric). Participants included 45 individuals (29 females, 16 males; 25 White, six Asian, six Other, five Black, two Latino, one biracial/multiracial individuals) taken from a convenience sample. Participants made their ratings on a 0–100 semantic differential scale with anchors labeled 'Very White/Eurocentric' to 'Very Black/Afrocentric.'

Objective measures and factor analysis Next, we measured a number of physical facial features. This allowed us to assess how various physical characteristics of the faces co-vary in creating meaningful psychological constructs. Based on a review of the social perception literature (Blair & Judd, 2010; Zebrowitz, 1997), we measured the median luminance of the face, nose width, nose length, lip thickness, face length, height and width of each eye, face width at the most prominent part of the cheek, face width at mouth, forehead length, distance between each pupil and the top of the head, distance between each pupil and the upper lip, chin length, length of cheek to chin for both sides of the face, and distance between pupils (see Table 1 for a list of all assessed facial features and a description of how each measure was derived). Before measuring, a guide was created so that research assistants could see how each measure was to be completed (this guide is posted on the CFD website for reference). Two research assistants independently completed the measures using Adobe Photoshop software. Once both raters finished measuring all of the faces, an absolute difference between the two measurements was computed. Differences greater than 20 % of the average were flagged and discussed by the research assistants. These differences were then reconciled and a final set of measures was obtained based on the raters' average. The inter-rater reliability of the physical measures was high ($rs \geq .74$).

Results

Subjective ratings

Analyses for this study were strictly descriptive. Our goals were to estimate the reliability of ratings and assess average raters' judgments of targets in the stimulus set. Because we had large amounts of randomly missing data in the subjective measures due to our sampling procedure, we calculated reliability using an estimation of interdependence procedure prescribed by Kenny and Judd (1996; see also Judd & McClelland, 1998). This technique yielded estimates of the reliability of single items, which were then submitted to the Spearman-Brown Prophecy Formula. Reliability for each judgment is presented in Table 2. Overall, reliabilities were high, ranging from .89 to .99. We caution, however, that these estimates are very likely to be inflated due to the large sample size of raters. Interrater reliability for the Eurocentricity-Afrocentricity ratings was also very high ($\alpha = .99$).

A number of significant correlations among subjective ratings emerged. In the interest of space, these correlations are presented in Table 3.

Objective measures and factor analysis

We submitted median face luminance, face length, face width at cheeks, face width at mouth, face shape, heartshapedness, nose shape, lipfullness, eye shape, eye size, upper head length, midface length, chin length, forehead height, cheekbone height, cheekbone prominence face roundness, and facial width to height ratio to an exploratory factor analysis using a principal component analysis with varimax rotation. A four-factor solution explained 72.26 % of the variance. Factor 1 corresponded with the facial width and had an Eigenvalue of 3.62 (20.11 % of the variance). Face width at cheeks, face width at mouth, face roundness, and a larger facial width to height ratio all loaded positively on Factor 1. Factor 2 corresponded with gender. Factor 2 had an Eigenvalue of 3.58 and explained 19.90 % of the variance among the variables. Cheekbone prominence, heartshapedness, eye shape, and eye size all positively loaded on Factor 2, while face length and chin length negatively loaded on Factor 2. Factor 3 clearly represented race and had an Eigenvalue of 3.21, explaining 17.81 % of the variance. Median luminance, chin length, and forehead height positively loaded on Factor 3, whereas nose shape and lip fullness negatively loaded on Factor 3. Finally, factor 4 reflected the upper to lower length ratio of the face. Factor 4 had an Eigenvalue of 2.60 and explained 14.45 % of the variance among the variables. Upper head length positively loaded on factor 4 and midface length, chin length, and cheekbone height negatively loaded on factor 4.

To better understand the factors resulting from the analysis of the principal components and to vet our labeling of these factors, we correlated the four factor scores with the subjective ratings of the faces. In the interest of brevity and clarity of presentation, we highlight only a few of those relationships here, but refer readers to Table 4 for the complete correlation matrix. In addition to corresponding with facial width metrics

Table 1 Facial features and measurements (Study 1)

Facial Feature	Measurement
Median Luminance	Median luminance of the face without neck or hair
Nose Width	Distance between outside edge of the nose at widest point
Nose Length	Distance between nose tip and upper edge of eyes at nose tip center
Lip Thickness	Distance between top and bottom edge of lips at thickest point
Face Length	Distance between bottom of chin to edge of top of forehead/hairline
Eye Height	Distance between upper and lower inner eyelid at pupil center (Right and Left measured separately and averaged)
Eye Width	Distance between inner and outer corner of eye (Right and Left measured separately and averaged)
Face Width at Most Prominent Part of the Cheek	Distance between the outer edges of the cheek at most prominent point
Face Width at Mouth	Distance between outer edges of cheeks at mid-mouth
Forehead Length	Distance from center of top of forehead/hairline to the center between the eyes at pupils
Distance Between Pupils	Distance between the center of each pupil
Distance Between Pupil and Top of Face	Distance between pupil center to top of forehead/hairline (Right and Left measured separately and averaged)
Distance Between Pupil and Upper Lip	Distance between pupil center to top edge of lips (Right and Left measured separately and averaged)
Chin Length	Distance from bottom edge of lips to base of chin
Length of Cheek to Chin	Distance between midcheek to bottom of chin (Right and Left measured separately and averaged)
Midbrow to Hairline	Distance between middle eyebrow to top of forehead/hairline (Right and Left measured separately and averaged)
Facial Width-to-Height Ratio (fWHR)	(Distance between the outer edges of the cheek at most prominent point) ÷ (Distance between upperlip and brow)
Face shape	(Face Width at Most Prominent Part of the Cheek) ÷ (Face Length)
Heartshapeness	(Face Width at Most Prominent Part of the Cheek) ÷ (Face Width at Mouth)
Nose shape	(Nose width) ÷ (Nose Length)
Lip Fullness	(Lip thickness) ÷ (Face length)
Eye Shape	(Eye height) ÷ (Eye width)
Eye Size	(Eye height) ÷ (Face length)
Upper Head Length	(Forehead length) ÷ (Face length)
Midface Length	(Distance between pupil and upper lip averaged for right and left side) ÷ (Face length)
Chin Size	(Chin length) ÷ (Face length)
Forehead Height	(Midbrow to Hairline averaged for right and left side) ÷ (Face length)
Cheekbone Height	(Length of cheek to chin averaged for right and left side) ÷ (Face length)
Cheekbon eProminence	(Face width at most prominent part of the cheek – Face width at mouth) ÷ (Face Length)
Face Roundness	(Face width at mouth) ÷ (Face length)

Note. Some calculations were borrowed from Blair and Judd (2010; Table 1)

All measures were completed in Adobe Photoshop. Length measures taken in pixels

in the factor analysis, stimuli with higher scores on factor 1 were seen as more feminine (r = .20, p = .01) and less masculine (r = -.19, p = .02). This is consistent with our judgment of the resulting factor scores, as males tend to have wider faces than females (Carré & McCormick, 2008). It is worth noting, however, that these correlations are relatively small. Correlations between these subjective measures and factor 2 corroborated our designation of factor 2 as gender. Factor 2 positively correlated with femininity (r = .54, p < .001) and negatively with masculinity (r = -.56, p < .001). Factor 3 shared a very strong, negative correlation with subjective ratings of Afrocentricity (r = -.93, p < .001) and thus clearly signaled race, as suggested by the factor loadings. Finally, factor 4, which we believe represents a larger upper to lower face length, was significantly (but not overwhelmingly) correlated with race and gender. Higher values on this factor score positively related to femininity (r = .17, p = .03) and Afrocentricity (r = .17, p = .03), but negatively related to

Table 2 Reliability measures of neutrally expressed targets in the Chicago Face Database (n = 1,087; Study 1)

Item Rated	α
Age	0.896
Attractiveness	0.998
Babyfacedness	0.996
Emotional Expressiveness – Angry	0.997
Emotional Expressiveness – Disgust	0.995
Emotional Expressiveness – Fear	0.992
Emotional Expressiveness – Happy	0.997
Emotional Expressiveness – Sad	0.995
Emotional Expressiveness – Surprise	0.993
Emotional Expressiveness – Threat	0.995
Femininity	0.999
Masculinity	0.999
Afrocentricity	0.994
Trustworthiness	0.993
Unusual	0.991

masculinity (r = -.16, p = .05). Because we did not include subjective ratings of face shape and size (e.g., length, width, roundness, etc.), we cannot fully test the labeling of factors 1 and 4; however, the principal components analysis showed strong relationships between facial width metrics (factor 1) and face length proportions (factor 4).

Discussion

The goal of Study 1 was to establish the CFD by collecting images of human targets who varied by gender and race and to begin gathering data about these targets. 158 participants were photographed under standardized conditions making a variety of emotional facial expressions. Images were processed to standardize the size of the targets' core features and equate for warmth. Next, we collected subjective and objective ratings of these targets. To begin, targets were rated on a number of dimensions (e.g., attractiveness, age, threat, etc.) by a large sample of participants in order to obtain norming data for the targets. We observed high reliability across all of our measures. Subsequently, we assessed the physical properties of the targets' faces and submitted these measures to a factor analysis. Factor analysis produced four factors that accounted for roughly 72 % of the variance among the measurements. Analysis of these factors and correlations between factor scores and subjective ratings produced psychologically meaningful correspondence. Among these factors were face width, gender, race, and upper to lower face length ratio.

Table 3 Correlations of subjective target ratings (Study 1)

Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Age (1)	26.88	6.84														
Attractiveness (2)	3.15	0.73	.13													
Babyfacedness (3)	2.69	0.59	.30**	.17*												
Emotional Expressiveness – Angry (4)	2.62	0.62	.06	.34**	.34**											
Emotional Expressiveness – Disgust (5)	2.38	0.48	.05	.33**	.30**	.93**										
Emotional Expressiveness – Fear (6)	2.29	0.38	.04	.26**	.13†	.40**	.43**									
Emotional Expressiveness – Happy (7)	2.54	0.58	.11	.47**	.34**	.73**	.70**	.55**								
Emotional Expressiveness – Sad (8)	2.92	0.52	.01	.38**	.13†	.41**	.49**	.64**	.70**							
Emotional Expressiveness – Surprise (9)	1.95	0.35	.13†	.04	.07	.11	.14†	.54**	.01	.12						
Emotional Expressiveness – Threat (10)	2.40	0.51	.12	.42**	.38**	.86**	.75**	.30**	.60**	.30**	.09					
Femininity (11)	3.14	1.26	.06	.36**	.23**	.13†	.02	.03	.21**	.10	.19*	.45**				
Masculinity (12)	3.56	1.24	.08	.22**	.23**	.13	.01	.07	.13	.04	.18*	.45**	.97**			
Eurocentricity-Afrocentricity (13)	51.81	29.11	.12	.10	.13	.01	.02	.06	.11	.05	.08	.09	.08	.11		
Trustworthiness (14)	1.46	0.18	.07	.66**	.32**	.71**	.66**	.40**	.81**	.52**	.02	.74**	.34**	.24**	.10	
Unusual (15)	2.56	0.42	.07	.32**	.11	.34**	.34**	.14†	.22**	.16*	.09	.42**	.16*	.19*	.10	.38**

† $p \leq .10$. * $p \leq .05$. ** $p \leq .01$

Table 4 Correlations of factor scores and subjective target ratings (Study 1)

Variable	Factor 1 Facial Width	Factor 2 Gender	Factor 3 Race	Factor 4 Upper to Lower Face Length Ratio
Age	.07	.13†	.04	.03
Attractiveness	.15*	.14†	.07	.05
Babyfacedness	.13†	.14†	.15†	.10
Emotional Expressiveness – Angry	.08	.05	.01	.07
Emotional Expressiveness – Disgust	.15†	.01	.04	.02
Emotional Expressiveness – Fear	.02	.27**	.06	.02
Emotional Expressiveness – Happy	.00	.08	.04	.07
Emotional Expressiveness – Sad	.05	.10	.01	.04
Emotional Expressiveness – Surprise	.03	.22**	.06	.03
Emotional Expressiveness – Threat	.08	.22**	.07	.06
Femininity	.20**	.54**	.13	.17*
Masculinity	.19*	.56**	.15†	.16*
Eurocentricity-Afrocentricity	.07	.03	.93**	.17*
Trustworthiness	.01	.06	.06	.12
Unusual	.01	.06	.14†	.12

† $p \leq .10$. * $p \leq .05$. ** $p \leq .01$

A critical piece of our undertaking was to develop a method for sharing this resource with researchers. Through the website, researchers can download high resolution image files for all 158 targets making neutral, fearful/afraid, angry, happy (both open and closed mouth smiles), and neutral expressions along with all of the data we described in Study 1. Since the completion of the studies described above, we have expanded the database by adding almost 450 additional targets to the database, for a total of over 600 unique targets. These include 57 Asian females, 53 Asian males, 95 Black females, 108 Black males, 56 Latinas, 52 Latinos, 90 White females, and 94 White males. We have also added 20 biracial East Asian/White European faces and hope to add more faces that vary in terms of age. Neutral-expression image files for each of these new targets have been added to the set of available stimuli and we will continue to add emotional expressions and norming data for these targets as well. Our hope is that this resource will evolve over time to include additional stimuli and data from users as the stimuli are utilized by different researchers in a variety of paradigms.

Study 2: Expert ratings of the CFD

The goal of Study 2 was to examine the suitability of the targets for use in psychological research. To assess the quality of the images in the CFD, we surveyed researchers in the field of social psychology who currently conduct or who have conducted empirical research using face stimuli. These expert ratings were then used for two purposes. First, we wanted to determine whether the targets were generally deemed as viable stimuli for research purposes. We also wanted to incorporate data about the usability of the targets into the database, because we anticipated that it might be useful for potential users to know how experts in the field judged the stimuli. Second, we sought to test whether researchers systematically attended to particular target features when rating the suitability of stimuli.

Method

Participants

Twenty-six participants completed the survey (15 male; 11 female). Participants included 20 Whites, two Asians, two Blacks, and two who self-identified as Other. The sample reported being 43.28 years old on average (SD = 11.32). When asked how many years they had been conducting psychological research, the 22 participants who responded reported an average of 21.09 years (SD = 10.81) and a combined 464 years of research experience.

Procedure

Researchers within the field of social psychology were contacted via email to participate in an online survey inviting them to view and rate the CFD targets. After providing consent, participants were randomly presented with targets block by target type (Black females, Black males, White females, and White males). Within each block type, targets were randomly presented. Participants

viewed targets one at a time and were asked to judge, "How suitable would the target pictured above be for a study requiring images of [Black females]?" on a 5-point Likert scale (1 = not at all suitable, 5 = highly suitable). Upon completing the survey, participants were invited to enter a draw for a $50 gift card or a donation to the charity of their choosing.

Results

Reliability analysis revealed high concordance among expert ratings (α = .99). Recall that the most important goals of the current study were to determine whether the targets were deemed suitable for research purposes and to incorporate data about the suitability of each target into the CFD database. We first computed average suitability ratings by collapsing across participants' ratings and compared this average to the midpoint of the ratings scale using a one-sample t-test. Overall, targets were judged as significantly more suitable for research than 'neutral,' $t(157) = 22.49$, $p < .001$ ($M = 3.95$, $SD = 0.53$). However, eight of the 158 targets received an average rating that fell below the midpoint of the scale. Next, we ran basic analyses to determine whether the targets differed in suitability by race and gender. In a target-level analysis, we regressed average suitability (deviated to the midpoint of the scale) on contrast-coded race (White = -1; Black = +1), contrast-coded gender (male = -1; female = +1), and the race × gender interaction. The effects of race, $t(154) = 0.28$, $p = .78$, gender, $t(154) = -0.13$, $p = .89$, and the race × gender interaction, $t(154) = 0.61$, $p = .54$ were not statistically significant. However, the intercept was significant, $t(154) = 22.11$, $p < .001$. These results reveal (a) no evidence that the race and gender categories differed in terms of how suitable they were judged and (b) that after accounting for the category memberships of the targets, the average stimulus was deemed suitable for research purposes (Fig. 2).

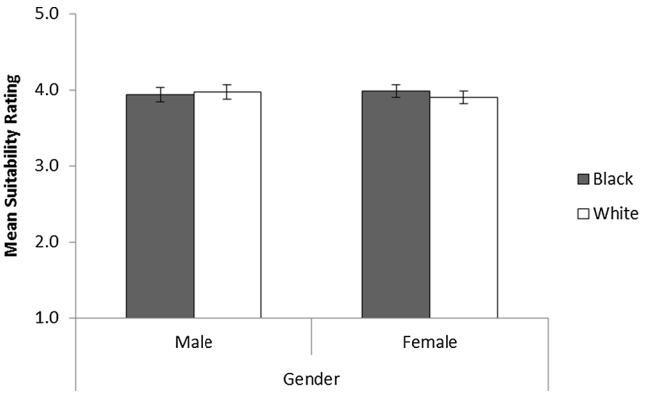

Fig. 2 Mean suitability ratings as a function of target race and gender (Study 2)

The second goal of the current study was to determine whether particular target features influenced participants' ratings of suitability. Given the item wording in the current study, which asked participants to think about how they would rate a given target if they needed stimuli involving a particular race and gender of targets, we predicted that participants would be strongly influenced by the racial and gender category fit of the targets. We used subjective ratings of Afrocentricity-Eurocentricity obtained in Study 1 to represent racial category fit (i.e., racial prototypicality). Recall that Afrocentricity was measured on a scale from 0 to 100. Racial prototypicality for Black targets was computed simply as a target's average Afrocentricity score; however in order to equate the scale for White targets, we took Afrocentricity scores for White targets and subtracted that value from 100. For example, a White target with an average Afrocentricity score of 17 yielded a racial prototypicality score of 83. We used Study 1 subjective ratings of masculinity and femininity to represent gender prototypicality for male and female targets, respectively. Next, taking target as our unit of analysis, we regressed suitability ratings on contrast-coded race (White = -1; Black = +1), contrast-coded gender (male = -1; female = +1), the race × gender interaction, mean-centered racial prototypicality, and mean-centered gender prototypicality (for reference, we will refer to this as the 'basic model' below). This allowed us to test the hypothesis that goodness of category fit influenced participants' ratings after accounting for the basic category structure. In this analysis, the effect of race was not significant, $t(152) = 0.07$, $p = .95$. However, a main effect of gender showed that, controlling for racial and gender prototypicality, female targets ($M = 4.07$, se = 044) were rated higher in suitability than male targets ($M = 3.80$, $se = .05$), $t(152) = 4.04$, $p < .001$. We also observed a race × gender interaction, $t(152) = 3.42$, $p = .001$. Within White targets there was no evidence for a difference between female ($M = 3.96$, $se = .06$) and male targets ($M = 3.91$, $se = .06$), $t(152) = 0.58$, $p = 56$. However, when holding prototypicality constant, Black females were judged as more suitable ($M = 4.18$, $se = .06$) than Black males ($M = 3.80$, $se = .07$), $t(152) = 3.42$, $p = .001$. Importantly, both prototypicality measures were also statistically significant. Racially prototypic targets were rated higher in terms of suitability, $t(152) = 7.39$, $p < .001$. Likewise, experts preferred targets that were higher in terms of gender prototypicality, $t(152) = 9.74$, $p < .001$. These data suggest that, across the four target categories, prototypicality was positively related to experts' judgments of suitability[1].

[1] We also tested two separate models in which 1) *racial* prototypicality and 2) *gender* prototypicality were allowed to interact with target category. Complete results are provided in the supplementary materials, which can be found on the CFD website.

Next, we examined how other subjective ratings of the targets related to ratings of suitability, controlling for the basic category structure and the effects of racial and gender prototypicality. For example, we wanted to test whether judgments of how unusual a face was rated by our norming sample related to judgments of suitability by experts. To do this, we augmented the basic model, adding mean-centered unusualness as a predictor. The effects of the basic-model predictors did not change. The critical question in this model was whether, over and above the basic model, subjective unusualness ratings related to experts' ratings of suitability. Analyses revealed that more unusual faces were indeed viewed as less suitable, $t(151) = -7.75$, $p < .001$. Identical analyses were conducted examining each of the subjective ratings in turn. Using this analytic strategy, we found that (over and above the basic model) subjective ratings in trust, $t(151) = 5.97$, $p < .001$, babyfacedness, $t(151) = 2.38$, $p = .02$, attractiveness, $t(151) = 4.24$, $p < .001$, and happiness, $t(151) = 4.28$, $p < .001$ were all significantly and positively related to suitability, whereas disgust, $t(151) = -5.80$, $p < .001$, threat, $t(151) = -7.03$, $p < .001$, surprise, $t(151) = -3.62$, $p < .001$, and fear, $t(151) = -3.44$, $p = .001$ were significantly and negatively related to suitability. Target differences in sadness, $t(151) = -1.20$, $p = .23$, and age, $t(151) = -0.28$, $p = .78$, did not reliably relate to suitability.

Discussion

Study 2 examined whether stimuli from the CFD were deemed suitable for research purposes by experts within the field of social psychology. We found that experts' ratings were highly reliable, despite having received no direction or basis on which to make their judgments. Overall, experts had a positive assessment of the CFD targets. Suitability averages and standard deviations for each target are included in the CFD documentation so potential users can access the data. Across all of the models we tested, mean suitability ratings, which were captured by the intercept, were consistently in the 3.8–3.9 range on a 5-point scale. We found that more racially and gender prototypic targets were viewed as more suitable, over and above the target's social category membership. Black targets who were more Afrocentric and White targets who were more Eurocentric, as judged by an independent group of lay participants, received higher suitability ratings. Likewise, experts preferred masculine males and feminine females to gender atypical targets. This was not altogether surprising given that participants were essentially asked to judge the face in terms of race and gender.

The high reliability among raters suggests that experts rely on similar aspects of target stimuli when making judgments. This suggests that, whether implicitly or explicitly, experts have common criteria in mind when judging the quality of stimuli. The question of *which* features affect raters was the second focus of Study 2. Here we found that targets who were judged by our norming sample to be displaying disgust, threat, surprise, and fear and targets who were rated as more unusual tended to receive lower suitability ratings. On the other hand, targets who were rated higher in trust, babyfacedness, attractiveness, and happiness were rated higher in terms of suitability. Several aspects of these findings are worth elaborating. Although most emotions detracted from suitability, targets who were rated as higher in happiness (within the small range of emotional variation displayed by these relatively neutral faces) actually received higher suitability ratings[2]. It is possible that the relationship between happiness and suitability is spurious, and actually reflects the strong correlation between happiness and attractiveness ratings, $r = .52$, $p < .001$, which also positively related to suitability. In order to test this, we ran an additional analysis in which we predicted suitability including both attractiveness and happiness scores in the same model. Here, we found that attractiveness, $t(150) = 2.99$, $p = .003$, and happiness, $t(150) = 3.05$, $p = .003$, independently predicted suitability controlling for race, gender, race × gender, racial prototypicality, and gender prototypicality.

Systematic preferences in the selection of research stimuli could have consequences for theory building. Preferences for particular characteristics or specific features may constitute a sampling bias, causing researchers to overlook important boundary conditions or inflate the size of an effect. Take for example the preference that experts showed for more prototypic targets. As we briefly alluded to in the "Introduction" Section, research in the area of stereotyping and prejudice (e.g., Blair et al., 2004; Livingston & Brewer, 2002; Ma & Correll, 2011) finds that individuals who possess more prototypic features are judged in accordance with group stereotypes and evaluations to a greater extent, and these findings occur even after accounting for social category membership (e.g., more prototypic Black males are rated more dangerous and less positively than their less prototypic counterparts). If researchers' preferences for faces predict actual selection and use of particular types of targets, they may inadvertently overestimate or engender particular effects in their studies. Essentially *any* preference for particular featural characteristics could bias the sample of stimuli. The consequence of those biases may be relatively minute or hugely consequential. Being cognizant of these influences and having access to highly variable stimulus sets with extensive pretesting data, such as the CFD, may help allay researchers detect and prevent some of these problems.

[2] As a reminder, experts were presented with images of targets making neutral facial expressions. It is possible that experts picked up on the same non-neutral aspects of the faces that raters did in Study 1 and/or that perceptions of faces as emotional occurred for other reasons (e.g. Hugenberg & Bodenhausen, 2003).

General discussion

Faces constitute a critical part of the social environment. This simple fact has translated to the extensive use of face stimuli in psychological research. Fortunately, researchers have many options for facial stimuli at their disposal, ranging from published databases (e.g., Ekman & Friesen, 1976; Lundqvist et al., 1998; Tottenham et al., 2009), to archives (e.g., Blair et al., 2004; Pauker & Ambady, 2009), to artificial faces (e.g., Matheson & McMullen, 2011), each of which has merits as well as drawbacks. This paper describes the development of a new database of facial stimuli and accompanying data including subjective ratings from a large sample and objective measures. The goal of this project was to create a useful, free resource for researchers needing a well controlled, high-resolution, demographically diverse stimulus set. Study 1 describes the initial development of the CFD, which includes 158 digital images of Black and White male and female targets, as well as the collection of norming data of subjective ratings of these targets and objective physical measures of each face. Digital photographs of each target were standardized in terms of the conditions under which photographs were taken as well as during post-production editing. Although the focus of the current paper is on the neutrally expressed images shot for each target, the CFD includes images of targets making a variety of emotional expressions. After editing and standardization, we obtained subjective ratings of each target along a number of socially relevant dimensions. Data were obtained from over 1,000 raters who showed a high degree of agreement in their assessments of the targets. Additionally, two independent raters took objective physical measures of each target face and these measures were submitted to a factor analysis to determine how these physical properties related with each other to create meaningful, latent constructs. Four factors emerged that represented face width, gender, race, and upper to lower face length ratio. Target images, subjective ratings, objective measurements, and factor scores were then aggregated and placed online so researchers could conveniently access the files. The focus of Study 2 was to assess experts' responses to the CFD targets and explore the types of characteristics researchers consider when selecting stimuli. Psychological scientists were solicited to provide ratings of how suitable each target was for research purposes. Overall, experts judged that the stimuli were suitable for research purposes. Additionally, respondents were highly reliable with respect to their judgments. Then, using the subjective ratings obtained in Study 1, we examined whether experts showed preferences for particular target characteristics when assessing the suitability of a target. Experts showed a strong preference for prototypic targets, with respect to both race and gender. Controlling for this preference, we also found that expert preferences correlated with how trustworthy, babyfaced, attractive, and happy faces were rated.

Contributions of the CFD

The stimuli we developed and the accompanying norming data fill several important niches within the available set of resources. One of the distinguishing features of the CFD is the diversity of the available targets. The first phase of development only included self-identified White and Black male and female targets; however, we have subsequently added Latino, Asian, and a small number of biracial Asian/White European targets. Increasing the racial diversity of available databases makes a significant contribution to many areas of study. For example, within the domain of stereotyping and prejudice, the vast majority of research has focused on dynamics between Black and White individuals (Sadler, Correll, Park, & Judd, 2012). Making Latino and Asian targets available may enable or at least facilitate researchers in broadening their investigations. In addition to racial diversity, the CFD also offers a large number of targets within each racial category. Whereas many databases include a small handful of non-White targets (e.g., Tottenham et al., 2009), the CFD has at least 40 unique targets within each race and gender category. This is an important feature of the CFD because it affords researchers the chance to carefully select stimuli and even systematically vary specific target characteristics within these social categories. This point highlights perhaps the most significant niche that the CFD fills within the existing database options. The CFD is composed of not just digital image files, but also offers fairly extensive data about each target. Researchers can use the available data on the targets to make their selections, saving resources that would otherwise be spent pre-testing stimuli. This provides both theoretical and practical applications for users.

We view this project as critically important and timely given recent concerns about the integrity of psychological research (Pashler & Wagenmakers, 2012) and the call for greater transparency in scientific reporting (Nosek & Bar-Anan, 2012). The development of a resource like the CFD has the potential to facilitate greater transparency by allowing others to view the exact stimuli that a researcher has used in a given study. This information could help consumers of research be more informed when evaluating research. The CFD will also make it easier to conduct exact replications, because researchers can use the same stimuli employed by other researchers (but see Stroebe & Strack, 2014). Availability of the data in the CFD can also help promote conceptual replications. In cases where researchers use different stimulus sets, data about target stimuli can be compared, which could prove informative for identifying boundary conditions, confounds, and the like.

Accessing the CFD and maintenance plan

As we have alluded to above, the CFD (i.e., image files and associated data) can be accessed at http://www.chicagofaces.org/. Researchers are asked to agree to the terms of use

indicated on the website and can download the entire database for free. We anticipate that the CFD will expand and evolve over time. Since the first phase of the development, we have expanded the database to over 600 unique Asian, Black, Latino, White, and biracial male and female targets. Subjective ratings and objective measures are currently being gathered for the newly added targets, however, target image files are already available online and data will be uploaded once data collection is completed. Moving forward, our maintenance plan for the CFD involves the continued accumulation of target data. Our hope is that researchers who use the CFD will collaborate in these maintenance efforts by reporting how targets perform in different experimental contexts. For example, given a suitable methodology, researchers who use the stimuli to study racial bias in an evaluative priming task might provide mean evaluation scores for each target, which could be added to our data. This repository of data may prove useful for others as they select stimuli or even in the development of novel research pursuits.

Acknowledgments Funding for this research was provided by the National Science Foundation (#1226143) to the first author and the University of Chicago Booth School of Business to the third author. We would like to thank Charles Judd and Bernadette Park for their consultation regarding analyses. We also extend our thanks to Kolina Koltai, Rebecca White, Megan Davis, and Michelle Revels.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., … Schaller, M. (2006). They all look the same to me (unless they're angry): From out-group homogeneity to out-group hetereogeneity. *Psychological Science, 17*(10), 836–840.

Asendorpf, J. B., Conner, M., de Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry, 38*(7), 813–822.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241–251.

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19,* 711–720.

Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y. W., … Forsyth, D. A. (2004). Names and faces in the news. *IEEE Conference on Computer Vision and Pattern Recognition, 2,* 848–854.

Blair, I. V., & Judd, C. M. (2010). Afrocentric facial features and stereotyping. In R. B. Adams Jr., N. Ambady, K. Nakayama, & S. Shimojo (Eds.), *The science of social vision*. New York: Oxford University Press.

Blair, I. V., Judd, C. M., & Fallman, J. L. (2004). The automaticity of race and Afrocentric facial features in social judgments. *Journal of Personality and Social Psychology, 87*(6), 763–778.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327.

Carré, J. M., & McCormick, C. M. (2008). In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B: Biological Sciences, 275,* 2651–2656.

Coin, C., & Tiberghien, G. (1997). Encoding activity and face recognition. *Memory, 5,* 545–568.

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

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027.

Fiedler, K. (2011). Voodoo correlations are everywhere – Not only in social neurosciences. *Perspectives on Psychological Science, 6,* 163–171.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14,* 640–643.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 180–232). New York, NY: McGraw-Hill.

Judd, C. M., Westerfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social pyshcology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69.

Kahneman, D. (2012, September 26). A proposal to deal with questions about priming effects.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302–4311.

Kenny, D. A., & Judd, C. M. (1996). A general procedure for the estimation of interdependence. *Psychological Bulletin, 119,* 138.

Kumar, N., Belhumeur, P. N., & Nayar, S. K. (2008). FaceTracer: A search engine for large collections of images with faces. *Proceedings of the 10th European Conference on Computer Vision (ECCV),* 340–353.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. *Attention and Orienting: Sensory and Motivational Processes,* 97–135.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming?: Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology, 82,* 5–18.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional faces*. Stockholm, Sweden: Karolinska Institute.

Ma, D. S., & Correll, J. (2011). Target prototypicality moderates racial bias in the decision to shoot. *Journal of Experimental Social Psychology, 47,* 391–396.

Ma, D. S., & Devos, T. (2013). Every heart beats true, for the red, white, and blue: National identity predicts voter support. *Analyses of Social Issues and Public Policy.*

Matheson, H. E., & McMullen, P. A. (2011). A computer-generated face database with ratings on realism, masculinity, race, and stereotypy. *Behavior Research Methods, 43,* 224–228.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36,* 630–633.

Morris, J. S., Friston, K. J., Büchel, C., Frith, C. D., Young, A. W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain, 121,* 47–57.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: A two-process theory of infant face recognition. *Psychological Review, 98,* 164.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23,* 217–243.

Pascalis, O., & Kelly, D. J. (2009). The origins of face processing in humans: Phylogeny and ontogeny. *Perspectives on Psychological Science, 4,* 200–209.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530.

Pauker, K., & Ambady, N. (2009). Multiracial faces: How categorization affects memory at the boundaries of race. *Journal of Social Issues, 65,* 69–86.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition, 24,* 607–640.

Ratner, K., Kaul, C., & Van Bavel, J. (2012). Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive and Affective Neuroscience, 8,* 750–755.

Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review, 5,* 659–669.

Rowley, H. A., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20,* 23–38.

Sadler, M. S., Correll, J., Park, B., & Judd, C. M. (2012). The world is not black and white: Racial bias in the decision to shoot in a multiethnic context. *Journal of Social Issues, 68,* 286–313.

Sato, W., & Yoshikawa, S. (2013). Recognition memory for faces and scenes. *The Journal of General Psychology, 140,* 1–15.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing A positron emission tomography study. *Brain, 115*(1), 15–36.

Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology, 36,* 763–798.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59–71.

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition, 13,* 657–665.

Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing, 28,* 902–913.

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

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., … Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research, 168,* 242–249.

Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research, 36,* 1761–1771.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115–1125.

Williams, K. D., & Jarvis, B. (2006). Cyberball: A program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods, 38,* 174–180.

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

Yong, E. (2012). Bad copy. *Nature, 485,* 298–300.

Zebrowitz, L.A. (1997) Reading faces: Window to the soul? Westview Press.