HHS Public Access

Author manuscript

Psychiatry Res. Author manuscript; available in PMC 2019 April 03.

Published in final edited form as:

Psychiatry Res. 2018 December ; 270: 1059–1067. doi:10.1016/j.psychres.2018.04.066.

The Racially Diverse Affective Expression (RADIATE) Face Stimulus Set

May I. Conley[a,*], **Danielle V. Dellarco**[b,*], **Estee Rubien-Thomas**[a], **Alexandra O. Cohen**[c], **Alessandra Cervera**[a], **Nim Tottenham**[d], and **BJ Casey**[a]

[a]Department of Psychology, Yale University, New Haven, Connecticut, USA

[b]Department of Psychology, University of Miami, Miami, Florida, USA

[c]Department of Psychology and Neural Science, New York University, New York, New York, USA

[d]Department of Psychology, Columbia University, New York, New York, USA

Abstract

Faces are often used in psychological and neuroimaging research to assess perceptual and emotional processes. Most available stimulus sets, however, represent minimal diversity in both race and ethnicity, which may confound understanding of these processes in diverse/racially heterogeneous samples. Having a diverse stimulus set of faces and emotional expressions could mitigate these biases and may also be useful in research that specifically examines the effects of race and ethnicity on perceptual, emotional and social processes. The racially diverse affective expression (RADIATE) face stimulus set is designed to provide an open-access set of 1,721 facial expressions of Black, White, Hispanic and Asian adult models. Moreover, the diversity of this stimulus set reflects census data showing a change in demographics in the United States from a white majority to a nonwhite majority by 2020. Psychometric results are provided describing the initial validity and reliability of the stimuli based on judgments of the emotional expressions.

Keywords

Emotion; Multiracial; Facial Expressions; Reliability; Stimuli; Validity

1. Introduction

Census data (2010, 2014) show a changing racial and ethnic distribution in the United States (U.S.) from a White, non-Hispanic Caucasian majority to a non-White majority (Colby and Ortman, 2014). It is predicted that by 2020, more than half of the children in the U.S. will be part of a non-White race or ethnic group. Since faces play a major role in communicating emotional and social information, facial expressions are often used as stimuli in psychological research. The number of studies published on face processing has more than

Corresponding Author: May I. Conley, Yale University, Department of Psychology, 2 Hillhouse Avenue, New Haven, CT 06511, 203-432-8444 (voice), 917-825-6729 (cell), May.Conley@Yale.edu.

*Authors contributed equally

Declaration of interest: none

tripled over the past decade from 5,000 (Tottenham et al., 2009) to over 15,000 based on a simple PubMed (2017) search of the terms: “face perception”, “face processing”, and “face expression”. Yet, psychological experiments presenting emotional faces often use stimuli from predominantly (or all) White face stimulus sets (e.g., Ekman and Friesen, 1976; Ekman and Friesen, 1978; Russell and Bullock, 1985; Lindquist et al., 1998). A growing literature of empirical studies shows that stimuli that portray individuals from minority and ethnic groups can have profound effects on perceptions and actions (Eberhardt et al., 2004; Funk et al., 2016). Face stimuli that reflect the racial and ethnic demographics of study participants are often needed for studies that use faces to examine these perceptual, emotional, and social processes. Diverse stimuli also afford testing of racial minority-majority effects on psychological processes. This paper introduces the open-access Racially Diverse Affective Expression (RADIATE) Face Stimulus Set (http://fablab.yale.edu/page/assays-tools) from over 100 racially and ethnically diverse adult models presented with initial validity and reliability scores.

Because facial expressions play a central role in communicating emotional and social information, they have been used broadly in psychological experiments. Several factors must be considered when selecting a stimulus set of emotional faces including model attributes (age, gender, race), total number of stimuli available, overall quality of the stimuli (resolution and image clarity, lighting, perceptual differences across models, etc.), fees or restrictions in use of the stimuli, and the range of emotions portrayed in the stimuli.

Initial stimulus sets of facial expressions were developed to examine processing and production of emotions (Ekman and Friesen, 1976; Ekman and Friesen, 1978; Russell and Bullock, 1985), however many of the earliest sets pose some potential limitations for researchers due to the narrow number of emotions depicted, number of models, or homogeneity of race and ethnicity (Erwin et al., 1992; Phillips et al., 1998; Winston et al., 2002). To address some of these limitations researchers have continued to create and standardize face stimuli (Hart et al., 2000; Kanade et al., 2000; Phelps et al., 2000; Gur et al., 2002; Batty and Taylor, 2003; Phelps et al., 2003; Tanaka et al., 2004; Pantic et al., 2005; Ashraf et al., 2009; Samuelsson et al., 2012; Dalrymple et al., 2013; Giuliani et al., 2013).

The growing use of face stimuli in psychological research has led to the creation of many stimulus sets, each with its own specific attributes (e.g., expanded set of models, ages, photographed angles, emotions, races and ethnicities). One of the largest available is the Karolinska Directed Emotional Faces (KDEF) Database (Lundqvist et al., 1998) which consists of 70 adult models each posing 7 different emotional expressions (Angry, Fearful, Disgusted, Sad, Happy, Surprised, and Neutral), photographed from 5 different angles, yielding a robust number of unique images (n= 4,900). However, only a subgroup of these stimuli has been validated (n= 490) (Goeleven et al., 2008) and all models are of European descent. Another large database, the Tarr Lab face database (Righi et al., 2012; Tarr, 2013) provides researchers with stimuli from over 200 Asian, African-American, Caucasian, Hispanic and Multiracial adult models. This large database is available in standard resolution movie stills or movie (mpeg) format and provides 3 different angles and 8 different emotions of mostly Caucasian or Asian models (68%).

More recently, several large stimulus sets have been developed that include child faces. The Radboud Faces Database (Langner et al., 2010) contains 1,176 faces of children and adults, the Dartmouth Database of Children's Faces (Dalrymple et al., 2013) contains 3,200 images of children's faces (aged 6–16), and the Child Affective Facial Expression (CAFE) set (LoBue and Thrasher, 2015) includes 1,192 emotional expressions from a highly diverse sample of child models (50% non-White stimuli, aged 2–8). Uniquely, the University of Oregon Emotional Expression Stimulus (DuckEES) set provides dynamic, moving stimuli of children and adolescents ages 8–18 (Giuliani et al., 2013). An advantage of these datasets are the inclusion of child models, however like KDEF, these databases include only or majority Caucasian models. The most recently published Developmental Emotional Faces Stimulus Set (DEFSS) similarly includes child models as well as adult models, and has some diversity, but the majority of the stimuli (87%) feature Caucasian models (Meuwissen et al., 2017).

Diversity in face stimuli is important for experiments that involve research participants from minority and ethnic groups that are growing in the U.S. (Colby and Ortman, 2014) because stimuli that portray individuals from racial in- versus out-groups can have significant effects on psychological processes and actions (Hart et al., 2000; Phelps et al., 2000; Golby et al., 2001; Elfenbein and Ambady, 2002; Lieberman et al., 2005; Herrmann et al., 2007; Rubien-Thomas et al., 2016). Face stimuli with diversity in both race and ethnicity are imperative for teasing apart perceptual, cognitive, and social processes and avoiding confounds due to using predominantly in- or all out-group stimuli.

Stimulus sets that include representation of non-Caucasian adult models have been generated, but many of these sets are still predominantly homogenous with limited diversity (Ekman and Matsumoto, 1993–2004; Mandal, 1987; Wang and Markham, 1999; Beaupre and Hess, 2005; Mandal et al., 2001). The Chicago Face Database (CFD), however, has a large set of standardized racially and ethnically diverse adult stimuli with widespread norming data for neutral expressions (e.g. face size, pupil size, attractiveness), but the number of expressions available for each model varies (Ma et al., 2015). Another database, the NimStim set of adult facial expressions, provides images of 43 models posing 16 expressions that include open- and closed-mouth variants of happy, sad, angry, fear, disgust, surprise, calm, and neutral faces (n=672). This sample includes approximately 40% non-White faces (10 African-American, 6 Asian-American, and 2 Latino-American models) in addition to validity and reliability scores for all 16 emotional expressions (Tottenham et al., 2009) laying the foundation for the current RADIATE face stimulus set.

The RADIATE face stimulus set consists of over 1,700 unique photographs of over 100 racially and ethnically diverse models (25% non-Hispanic White and 75% minority or ethnic group). Each model posed 16 different facial expressions (Tottenham et al., 2009) providing a wide range of emotions in a racially and ethnically diverse stimulus set. Images from the RADIATE stimulus set were developed so that they could be combined with other face stimulus sets with a standardized scarf-template that is included with the face stimuli (see Supplemental Text). The inclusion of the scarf-template affords researchers more flexibility in the selection and number of racially diverse stimuli by providing a standardized tool for harmonizing stimuli from other databases. Thus, the RADIATE stimulus set overcomes

potential limitations within pre-existing stimulus packages, by providing a large, standardized set of 16 emotional expressions by racially and ethnically diverse models to help move the field forward in the consideration of race and ethnicity effects on psychological processes.

2. Methods

2.1 Participants

Participants (*n*=693) within the United States were recruited using Amazon's Mechanical Turk (MTurk). All data from 31 participants were removed due to incomplete surveys or having duplicate IP addresses in our attempt to avoid repeat ratings from the same MTurk participants on a given survey. The final sample included 662 participants (260 female, 402 males) with a mean age of 27.6 (Range= 18–35, SD = 3.8) and self-identified as Asian (*n*=48), Black/ African American (*n*=70), Caucasian (*n*=470), Hispanic (*n*=63), and Other or Mixed races (*n*=11). All participants provided consent approved by the institutional review board at Yale University and were paid for taking part in the experiment.

2.2 Stimuli

A collection of 1,721 unique photographs of 109 adults (56 female, 53 male; 18–30 years old) posing 8 expressions (angry, calm, disgust, fear, happy, neutral, sad, and surprise) with open- and closed- mouth variants was obtained from the New York City metropolitan area in 2016 and harmonized and subsequently rated in 2016–2017. Twenty-three expressions were not captured in instances where models were unable to pose less-common expressions (e.g. sad open-mouth). Models were recruited from a community sample and were Asian (*n*=22), Black/African American (*n*= 38), Caucasian (*n*=28), Hispanic or Latino (*n*=20) and Other (*n*=1). All models consented to be photographed and released their photos to be used for research and scientific purposes, and all models were paid for their time.

Models were trained to make eight emotional expressions with open- and closed- mouth variants, given examples of each expression, and time to practice posing in a mirror before being photographed. See Supplemental Table 1 for instructions given to models for posing each facial expression. Models provided open- and closed-mouth variants of all expressions except the emotion of surprise for which only an open-mouthed expression was photographed. In addition, three versions of happy (closed-mouth, open-mouth, and high arousal open-mouth/exuberant) were obtained, based on prior work (Tottenham et al., 2009). Prior to being photographed, models were asked to remove any accessories that would visually separate them from the other models (i.e. glasses, headbands, hats). Models were photographed against a white wall and draped with a white scarf to hide clothing and reduce any potential reflected hues on the models' faces. Adobe Photoshop was used to correct for differences in luminosity, head size, and head position (See Supplemental Text for additional details on standardization of stimuli). The scarf mask and the stimuli are available in both black and white and color as supplemental materials and at http://fablab.yale.edu/page/assays-tools. [1]

2.3 Stimulus rating procedure

After editing and standardizing the images, the stimuli were uploaded to Qualtrics to be rated by MTurk participants. Surveys were subdivided into MTurk surveys[2] that contained subsets (emotional expressions from 8 to 9 models) of all 1,721 images to keep rating time to 20–30 minutes to be consistent with other facial stimuli rating tasks hosted in MTurk (Ma et al., 2015). Each stimulus was rated by an average of 50.92 MTurk participants (S.D.= 1.26, Range= 49–53) and detailed demographic information for each survey can be found in Supplemental Table 5. Each MTurk task contained links to 3 Qualtrics surveys. The first survey contained images of all expressions posed by 8–9 models presented in a randomized order and instructed participants to select the emotion (angry, calm, disgust, fear, happy, neutral, sad, surprise, or "none of the above") being depicted in the image. To prevent subjects from immediately re-rating stimuli for re-test reliability, the second survey contained a simple cognitive task, which asked participants to list as many words as possible starting with a given letter or within a given object-category (e.g. vegetable, animal). Participants had a minute for each category and a total of 3 letters and 3 objects were given to each participant. Upon completion of the cognitive task, the third survey was administered, presenting the exact same images from the first survey in a new randomized order. The first and third surveys were self-paced. After each survey, participants were given a code that was pasted into a box to confirm completion of all surveys and paid.

2.3.1 Validity ratings—Initial validity ratings were obtained in the first of the two surveys by having participants select one of 8 emotion-labels (angry, calm, disgust, fear, happy, neutral, sad, surprise) for each presented face or given the option to choose "none of the above". The "none of the above" choice was included to avoid inflation in correct labeling inherent in strict forced choice designs. Stimuli were presented in a randomized order that was counterbalanced across participants.

2.3.2 Reliability—Initial reliability ratings were calculated from all participants who were asked to rate the same subset of the 1,721 face stimuli a second time. Stimuli were presented in a new randomized order for each participant.

2.5 Data analytic procedures

2.5.1 Validity—Two validity measures (proportion correct and Cohen's kappa (Cohen, 1960)) were calculated for each of the 1,721 stimuli modeled after the analyses of the NimStim Set of Facial Expressions (Tottenham et al., 2009). Proportion correct was calculated by comparing the number of participants who correctly endorsed target expressions to the total number of ratings for each stimulus. Though proportion correct is often reported in examinations of facial expression (Ekman and Friesen, 1976; Mandal, 1987; Biehl et al., 1997; Wang and Markham, 1999; Mandal et al., 2001; Beaupre and Hess, 2005), Tottenham (2009) suggests that Cohen's kappa (Cohen, 1960) may be a better

[1]By downloading the RADIATE stimuli, users are agreeing to use the stimuli solely for approved institutional research or educational purposes and will not use them in any way to deliberately or inadvertently identify the individuals in the pictures.

[2]Thirteen MTurk surveys were initially administered, however an additional 2 models (BM18 and AM10) were rated separately with 11 individual stimuli that had not been processed with the first 13 surveys. Data from survey 14 was processed with the primary 13 ratings and these data are provided.

dependent variable for evaluations of stimulus sets since proportion correct does not consider false positives (Erwin et al., 1992). Kappa scores, a measure of agreement between participants' labels and models' expressions adjusted for agreement due to chance, were used to estimate agreement between selected labels and intended expressions. These scores were calculated across models within each survey, independently for open- and closed-mouth conditions (see Supplemental Text for detailed information on this calculation).

Endorsements of "none of the above" were counted as incorrect. Because of the nuanced differences between the calm and neutral expressions (see Figure 1 for expression examples and Supplemental Table 1 for instructions), ratings of "calm" and "neutral" were counted as correct for both calm and neutral expressions (Tottenham et al., 2009).

Additionally, a confusion matrix showing mean proportion of MTurk participants endorsing each expression (targets and non-targets) was used to further examine the breakdown of endorsements across expressions. This matrix reiterates proportion correct scores described previously in conjunction with proportion incorrect scores, revealing potential trends in errors (i.e. some emotions were consistently mistaken for the same expressions).

2.5.2 Reliability—Reliability was measured by calculating proportion agreement within subjects between the first and second ratings of each of the 1,721 face stimuli. Ratings from the second survey were not included in computing validity ratings.

3. Results

3.1 Validity

Validity ratings for each of the 16 emotional expressions are presented in Table 1 and Figure 2 (See Supplemental Tables 2a and 2b for individual proportion correct and kappa scores and Supplemental Table 3 for each actor). The overall proportion correct was high (Mean= 0.71, S.D.= 0.19, Median= 0.75). Kappa scores (the overall measure of agreement between participants' labels and models' intended expressions, adjusting for agreement due to chance), were also substantial (Mean= 0.65, S.D.= 0.19, Median= 0.68) indicating stimuli accurately conveyed their intended expressions (Landis and Koch, 1977).

Kappa scores per actor ranged from 0.6–0.8 in seventy-five percent of the models (i.e., 82 of the total 109 models) reflecting general agreement between participants' labels and models' expressions, adjusting for agreement due to chance (Cohen, 1960; Landis and Koch 1977) and fifty-four percent (59/109) of the mean proportion correct scores were above 0.70. Nearly seventy percent (11/16) of the mean kappa scores calculated for each expression were substantial (Landis and Koch, 1977) including the following expressions: angry-closed, angry-open, calm-closed, calm-open, disgust-open, happy-closed, happy-open, happy-exuberant, neutral-open, neutral-closed, and sad-closed. Mean proportion correct scores were high across expressions, with more than half (9/16) above 0.7 including: calm-closed, calm-open, disgust-open, happy-closed, happy-open, happy-exuberant, neutral-closed, neutral-open, and surprise.

The confusion matrix (Table 2) depicts average proportion of target and non-target labels endorsed for each expression. This matrix demonstrates that expressions were rarely identified as "none of the above" (endorsement across expressions ranged from 0.00–0.05) and that particular poses were consistently mistaken for other expressions. Table 1 and Figure 2 show low initial validity (proportion correct and kappa) scores for the fear-closed and fear-open expressions, which were consistently mislabeled as surprise (mean fear-closed labeled as surprise= 0.43, S.D.= 0.27, mean fear-open labeled as surprise= 0.41, S.D.= 0.24). Similarly, the open variant of the sad expression, an expression that many models had difficulty posing, was consistently mislabeled as disgust (mean sad labeled as disgust= 0.25, S.D.= 0.22).

3.2 Reliability

Reliability scores (i.e. proportion agreement) for each emotional expression are reported in Table 3 and presented in Figure 3, and reliability values for individual stimuli can be found in Supplemental Table 4. Overall, there was consistency between the first and second ratings with a mean reliability of 0.70 (S.D.= 0.16).

Approximately half (7/16) of the mean reliability scores for each expression ranged from 0.71–0.96 and the remaining (9/16) expressions ranged from 0.52–0.7. Sad-open ($Mean_{agreement}$= 0.52, S.D.=0.13) and calm-open ($Mean_{agreement}$= 0.58, S.D.= 0.11) faces had the greatest variability between rating sessions.

3.3 Race and Ethnicity of Model

Exploratory examination of variability in accuracy by race of model was performed. Validity and reliability scores are provided for each individual model for each expression in Supplemental Tables 2–4. Figure 4 provides mean proportion correct for ratings of emotion by race of the model illustrating overall moderate consistency in ratings for each race, however further studies optimized to explore the relationship between emotion perception and race of target are warranted. The variability observed across emotional categories by race of models is primarily for negative emotions. Regardless, mean accuracy for emotional categories for each race was high (70% or above) for 10 of the 16 emotions and above 50% for 3 emotional categories. The remaining expressions of fear-closed mouth, fear-open mouth and sad-open mouth were consistently low across all 4 races (below 50% accuracy). Most of the participants in the current sample were White (470 of 662). Nonetheless, preliminary accuracy plots are provided for each major racial group of participants for visualization purposes in Supplemental Figures 1a–1d. Further research will be needed to assess any effects of participant and model race on the accuracy of emotional ratings.

4. Discussion

This article presents the RADIATE Face Stimulus Set with initial validity and reliability scores calculated from judgments made by research participants on Amazon's MTurk. The large number of racially and ethnically diverse models posing a variety of facial expressions is provided as an open-access resource for researchers interested in psychological processes involving face processing and social stimuli.

Initial validity ratings were obtained with mean proportion correct of 0.71. Because proportion correct does not reflect false alarm judgments, kappa scores were calculated for each stimulus to measure agreement between participants' labels and models' intended expressions, adjusting for agreement due to chance. These calculations are based on the entire stimulus set. While some researchers report validity statistics for the stimuli with the highest scores (Dalrymple et al., 2013; Giuliani et al., 2013), other groups (e.g., Ma et al., 2015) do not report validity data and note the need for diverse stimuli to address demographic homogeneity across available databases. The comprehensive RADIATE stimulus set is provided because of the need for representative stimuli of individuals of color, especially considering the growing literature examining in- and out-group interactions (Byatt and Rhodes, 1998; Hart et al., 2000; Golby et al., 2001; Hugenberg et al., 2003; Herrmann et al., 2007; Lieberman et al., 2005). However, for researchers who are primarily interested in high emotion-identification accuracy, we provide Supplemental Tables 6a and 6b with percentage of models per race, sex, and emotion category above .70 to aid researchers in determining which category of RADIATE stimuli may be suitable for their purposes. In addition we provide proportion correct and Kappa scores for each of the 1,721 stimuli in Supplemental Tables 2a and 2b.

The validity scores vary across emotion categories, which is consistent with previous face stimulus sets (Ekman and Friesen, 1976; Tottenham et al., 2009) and is likely due to differences in emotion recognition across expressions (Strauss and Moscovitch, 1981; Calder et al., 2003). As previously shown, happy expressions typically have a high identification rate (Hare et al., 2005) while negative expressions have a lower identification rate (Biehl et al., 1997; Lenti et al., 1999; Calder et al., 2003; Calvo and Lundqvist, 2008; Palermo and Coltheart, 2004; Gur et al., 2002; Elfenbein and Ambady, 2003). For example, fear faces are often recognized less accurately than other expressions, particularly in forced-choice design studies (Russell, 1994) and sad faces are often mislabeled as neutral or disgusted (Palermo and Coltheart, 2004). As Tottenham and colleagues (2009) report, the open- and closed-mouth variants were produced to control for perceptual differences, however this manipulation may have precluded both production and recognition of certain expressions (e.g. some models had difficultly maintaining an expression while adjusting mouth-variant). Some of the less prototypical expressions (e.g. fear-closed, sad-open) are usually not included in face stimulus sets so the lower scores for these categories are not surprising.

It is important to consider the impact that context and culture may have on the perception of emotion in addressing the variance in emotion categories within the RADIATE face stimulus set. Although emotion perception is often considered static, research has demonstrated that contextual and cultural factors from a variety of sources (e.g. culture of stimulus and perceiver, emotion of a situation, or words) can affect how an emotional expression is perceived (Aviezer et al., 2008; Barrett et al., 2011; Kim et al., 2004). Therefore, we provide all stimuli and ratings to facilitate informed decisions regarding which stimuli are appropriate for the experimental design and objectives of future studies.

The exploratory plots illustrating accuracy of emotional categories for each race were moderately consistent. These findings are promising given that the majority of raters were

White participants, while the majority of the stimuli are of non-White models. However, future research will be needed to address associations among participant race, model race, and accuracy of categorizing emotional stimuli to support this claim.

Reliability ratings were calculated demonstrating overall consistency across both rating 1 and rating 2 with mean reliability of 0.70. However, there is variance across emotion categories, particularly for the sad-open and calm-open expressions. Given that reliability is a necessary condition for validity, some individual stimuli may not be appropriate for use in studies for which high reliability of emotional category is required.

An advantage of the RADIATE stimulus set is that images were rated using a semi-forced choice design, allowing participants to choose across 9 options (angry, calm, disgust, fear, happy, neutral, sad, surprised, or “none of the above”) for each expression. Consistent with the NimStim (Tottenham et al., 2009) methods, the “none of the above” choice was included because strict forced choice tasks can inflate correct labeling. However, Russell (1994) notes that the subtle complexities of expressions may not fully be captured with this design and that a combination of forced-choice, freely chosen, or spectrum (i.e. slightly happy, moderately happy, very happy) labels may be more appropriate for rating faces. Considering the number of stimuli in the RADIATE set, a major obstacle was obtaining expression ratings of all stimuli from a relatively large sample. To overcome this challenge we used Amazon’s MTurk, which is widely used within the research community (Buhrmester et al., 2011; Mason et al., 2012; Crump et al., 2013). However, using this approach required subdivided groups of models’ images, an approach used in other stimulus rating studies (Ma et al., 2015) and the inclusion of additional expression labels would have required further subdivision of models’ images.

The use of MTurk as the sole subject pool is a potential limitation of the study. While all subjects executed this study online, environmental factors such as background noise, room lighting, or the presence of other people could not be controlled and could account for the variance in total survey completion times (Mean= 18.12 minutes, S.D.= 12.83 minutes, Range= 5.34–198.17 minutes). Differences in timing to complete surveys also may be attributed to differences in Internet speed (e.g. image loading) or other individual differences (e.g. rating speed, leaving a browser window open). However, these factors are not known due to the nature of collection methods. Nonetheless, MTurk participants are often recruited to participate in psychology studies, some of which use face stimuli (Tran et al., 2017). The current participants were however “untrained” in that they were not subjected to using the Facial Action Coding System (FACS) (Ekman and Friesen, 1978) to evaluate expressions (Tottenham et al 2009).

MTurk is often used to obtain robust community samples that can be more diverse than college samples (Paolacci and Chandler, 2014; Mason and Suri, 2012). While research comparing MTurk participants to in-person samples is limited, one such study (Casler et al., 2013) showed that data obtained via MTurk, social media, and in-person did not vary across these different collection-methods. Additionally, Hauser and Schwarz (2016) found that MTurk participants were more attentive to task instructions than college students studied in-person. Further, in the current study, data were examined for incomplete surveys and

suspected duplicate-raters and removed prior to analysis to help minimize the impact of this data collection procedure on the quality of measures.

Another potential limitation of the current study is that most of the MTurk participants self-identified as Caucasian, while the majority of the stimuli were of minority and ethnic groups (75%), which has been demonstrated to interfere with face recognition (Levin, 1996; Byatt and Rhodes, 1998; Levin, 2000; MacLin and Malpass, 2001). Even so, we show substantial validity with Kappa scores per model ranging from 0.6–0.8 in seventy-five percent (82 of 109) of the models, reflecting general agreement between participants' labels and models' expressions (Cohen, 1960; Landis and Koch 1977).

In addition to the limitations present in the rater population, model attributes must also be considered. Models in the RADIATE set were from a community-sample opposed to a sample of trained models. Additionally, while the set contains some non-stereotypical depictions of gender (e.g. women with short hair, men with long hair), researchers interested in issues of gender-representation or gender-bias should consider including a greater number of androgynous models.

The main objective for developing the RADIATE Face Stimulus Set was to create a large, racially and ethnically diverse set of facial expressions that research participants could accurately identify. The stimulus set contains 1,721 images available in black and white and color and a standard scarf template offering researchers flexibility to combine RADIATE stimuli with other facial expression packages. This stimulus set is available to the scientific community as supplemental material and at http://fablab.yale.edu/page/assays-tools. Diverse stimulus sets of this nature may prove useful in examining psychological processes, especially considering the shift from majority to minority trends in the U.S., by providing representative stimuli that reflect the race and ethnicity of research participants and for testing questions specific to processing of in- versus out-group effects on psychological processes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Funding

This work was supported in part by a National Science Foundation Graduate Research Fellowships (A.O.C., E.R.-T.); a NIH U01 DA041174 grant (BJ.C.); a grant from the John D. and Catherine T. MacArthur Foundation to Vanderbilt University. Its contents reflect the views of the authors, and do not necessarily represent the official views of either the John D. and Catherine T. MacArthur Foundation or the MacArthur Foundation Research Network on Law and Neuroscience (www.lawneuro.org).

References

Adobe, 2005 Optimum strategies for using Adobe Photoshop CS2 in scientific and medical imaging https://www.adobe.com/digitalimag/pdfs/scimedwp.pdf. (accessed 28 August 2016).

Ashraf AB, Lucey S, Cohn JF, Chen T, Ambadar Z, Prkachin KM, Solomon PE, 2009 The painful face – pain expression recognition using active appearance models. Image and Vision Computing 27 (12), 1788–96. 10.1016/j.imavis.2009.05.007 [PubMed: 22837587]

Aviezer H, Hassin RR, Ryan J, Grady C, Susskind J, Anderson A, … Bentin S (2008). Angry, disgust, or afraid? Studies on the malleability of emotion perception. Psychological Science, 19(7), 724–732. 10.1111/j.1467-9280.2008.02148.x [PubMed: 18727789]

Barrett LF, Mesquita B, Gendron M (2011). Context in emotion perception. Current Directions in Psychological Science, 20(5), 286–290. 10.1177/0963721411422522

Batty M, Taylor MJ, 2003 Early processing of the six basic facial emotional expressions. Cogn. Brain Res 17 (3), 613–620. 10.1016/S0926-6410(03)00174-5

Beaupre MG, Hess U, 2005 Cross-cultural emotion recognition among Canadian ethnic groups. Journal of Cross-Cultural Psychology 36 (3), 355–370. 10.1177/0022022104273656

Biehl M, Matsumoto D, Ekman P, Hearn V, Heider K, Kudoh T, Ton V, 1997 Matsumoto and Ekman's Japanese and Caucasian facial expressions of emotion (JACFEE): reliability data and cross-national differences. Journal of Cross-Cultural Psychology 21 (1), 3–21. 10.1023/A:1024902500935

Buhrmester M, Kwang T, Gosling SD, 2011 Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6 (1), 3–5. 10.1177/1745691610393980 [PubMed: 26162106]

Byatt G, Rhodes G, 1998 Recognition of own-race and other-race caricatures: implications for models of face recognition. Vision Research 38 (15–16), 2455–2468. 10.1016/S0042-6989(97)00469-0 [PubMed: 9798009]

Calder AJ, Keane J, Manly T, Sprengelmeyer R, Scott S, Nimmo-Smith I, Young AW, 2003 Facial expression recognition across the adult life span. Neuropsychologia 41 (2), 195–202. 10.1016/S0028-3932(02)00149-5 [PubMed: 12459217]

Calvo MG, Lundqvist D, 2008 Facial expressions of emotion (KDEF): Identification under different display-duration conditions. Behavioral Research Methods 40 (1), 109–115. 10.3758/BRM.40.1.109

Casler K, Bickel L, Hackett E., 2013 Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. Computers in human behavior 29(6), 2156–2160.

Cohen J, 1960 A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1),37–46. 10.1177/001316446002000104

Colby SL, Ortman JM, 2014 Projections of the Size and Composition of the U.S. Population: 2014 to 2060, Current Population Reports, P25–1143, U.S. Census Bureau, Washington, D.C.

Crump MJC, McDonnell JV, Gureckis TM, 2013 Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS One 8 (3): e57410 10.1371/journal.pone.0057410 [PubMed: 23516406]

Dalrymple KA, Gomez J, Duchaine B, 2013 The Dartmouth Database of Children's Faces: Acquisition and validation of a new face stimulus set. PLoS ONE 8: e79131 10.1371/journal.pone.0079131 [PubMed: 24244434]

Eberhardt JL, Goff PA, Purdie VJ, Davies PG, 2004 Seeing black: race, crime, and visual processing. Journal of Personality and Social Psychology 87(6), 876–893. 10.1037/0022-3514.87.6.876 [PubMed: 15598112]

Ekman P, Friesen WV, 1976 Pictures of Facial Affect Consulting Psychologists Press; Palo Alto, CA.

Ekman P, Friesen WV, 1978 Facial Action Coding System: A Technique for the Measurement of Facial Movement Consulting Psychologists Press; Palo Alto, CA.

Ekman P, Matsumoto D Japanese and Caucasian Facial Expressions of Emotion (JACFEE) Consulting Psychologists Press; Palo Alto, CA: 1993–2004.

Elfenbein HA, Ambady N, 2002 On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin 128 (2), 203–235. 10.1037/0033-2909.128.2.203 [PubMed: 11931516]

Elfenbein HA, Ambady N, 2003 When familiarity breeds accuracy: cultural exposure and facial emotion recognition. Journal of Personality and Social Psychology 85 (2), 276–290. 10.1037/0022-3514.85.2.276 [PubMed: 12916570]

Erwin R, Gur RC, Gur RE, Skolnick B, Mawhinney-Hee M, Smailis J, 1992 Facial emotion discrimination: I. Task construction and behavioral findings in normal subjects. Psychiatry Research 42 (3), 231–240. 10.1016/0165-1781(92)90115-J [PubMed: 1496055]

Funk F, Walker M, Todorov A, 2016 Modeling perceptions of criminality and remorse from faces using a data-driven computational approach. Cognition and Emotion 31 (7), 1–13. 10.1080/02699931.2016.1227305

Giuliani NR, Flournoy JC, Ivie EJ, Von Hippel A, Pfeifer JH, 2017 Presentation and validation of the DuckEES child and adolescent dynamic facial expression stimulus set. International Journal of Methods in Psychiatric Research 26:e1553 10.1002/mpr.1553

Goeleven E, De Raedt R, Leyman L, Verschuere B, 2008 The Karolinska Directed Emotional Faces: a validation study. Cognition and Emotion 22 (6), 1094–1118. 10.1080/02699930701626582

Golby AJ, Gabrieli JD, Chiao JY, Eberhardt JL, 2001 Differential responses in the fusiform region to same-race and other-race faces. Nature Neuroscience 24 (8), 845–850. 10.1038/90565

Gur RC, Sara R, Hagendoorn M, Marom O, Hughett P, Macy L, Turner T, Bajcsy R, Posner A, Gur RE, 2002 A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. Journal of Neuroscience Methods 115 (2), 137–143. 10.1016/S0165-0270(02)00006-7 [PubMed: 11992665]

Hare TA, Tottenham N, Davidson MC, Glover GH, Casey BJ, 2005 Contributions of amygdala and striatal activity in emotion regulation. Biological Psychiatry 57 (6), 624–632. 10.1016/j.biopsych.2004.12.038 [PubMed: 15780849]

Hart AJ, Whalen PJ, Shin LM, McInerney SC, Fischer H, Rauch SL, 2000 Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. Neuroreport 11 (11), 2351–2355. [PubMed: 10943684]

Hauser DJ, Schwarz N, 2016 Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behav Res Methods 48 (1), 400–7. 10.3758/s13428-015-0578-z [PubMed: 25761395]

Herrmann MJ, Schreppel T, Jager D, Koehler S, Ehlis AC, Fallgatter AJ, 2007 The other-race effect for face perception: an event-related potential study. Journal of Neural Transmission 114 (7), 951–957. 10.1007/s00702-007-0624-9 [PubMed: 17318308]

Hope RM, 2013 Rmisc: ryan miscellaneous. R package version 1.5 https://CRAN.R-project.org/package=Rmisc

Hugenberg K, Bodenhausen GV, 2003 Facing prejudice: implicit prejudice and the perception of facial threat. Psychological Science 14 (6), 640–643. 10.1046/j.0956-7976.2003.psci_1478.x [PubMed: 14629699]

Kanade T, Cohn JF, Tian Y, 2000 Comprehensive database for facial expression analysis. IEEE Conference on Automatic Face and Gesture Recognition Grenoble, France.

Kim H, Somerville LH, Johnstone T, Polis S, Alexander AL, Shin LM, & Whalen PJ (2004). Contextual Modulation of Amygdala Responsivity to Surprised Faces. Journal of Cognitive Neuroscience, 16(10), 1730–1745. 10.1162/0898929042947865 [PubMed: 15701225]

Landis JR, Koch GG, 1977 The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174. [PubMed: 843571]

Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A, 2010 Presentation and validation of the Radboud Faces Database. Cognition and Emotion 24 (8), 1377–1388. 10.1080/02699930903485076

Lenti C, Lenti-Boero D, Giacobbe A, 1999 Decoding of emotional expressions in children and adolescents. Perceptual and Motor Skills 89 (3), 808–814. 10.2466/pms.1999.89.3.808 [PubMed: 10665012]

Levin DT 1996 Classifying faces by race: The structure of face categories. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 1364–82. http://dx.doi.org/10.1037S0096-3445.129.4.55

Levin DT, 2000 Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categoires and the cross-race recognition deficit. Journal of Experimental Psychology: General 129, 559–574. 10.1037/0096-3445.129.4.559 [PubMed: 11142869]

Lieberman MD, Hariri A, Jarcho JM, Eisenberger NI, Bookheimer SY, 2005 An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. Nature Neuroscience 8, 720–722. 10.1038/nn1465 [PubMed: 15880106]

LoBue V, Thrasher C, 2015 The Child Affective Facial Expression (CAFE) Set: Validity and reliability from untrained adults. Frontiers in Emotion Science 5 (1532), 1–8. 10.3389/fpsyg.2014.01532

Lundqvist D, Flykt A, Öhman A. Karolinska Directed Emotional Faces, KDEF (CD ROM). 1998 Stockholm, Sweden: Karolinska Institute, Department of Clinical Neuroscience, Psychology Section

Ma D, 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. 10.3758/s13428-014-0532-5 [PubMed: 25582810]

Mason W, Suri S 2012 Conducting behavioral research on Amazon's Mechanical Turk. Behavioral Research Methods 44 (1), 1–23. 10.3758/s13428-011-0124-6

MacLin OH, Malpass RS, 2001 Racial categorization of faces: The ambiguous race face effect. Psychology Public Policy, and Law 7 (1), 98–118. 10.1037/1076-8971.7.1.98

Mandal MK, 1987 Decoding of facial emotions, in terms of expressiveness, by schizophrenics and depressives. Psychiatry 50 (4), 371–376. 10.1080/00332747.1987.11024368 [PubMed: 3423162]

Mandal MK, Harizuka S, Bhushan B, Mishra RC, 2001 Cultural variation in hemifacial asymmetry of emotion expressions. British Journal of Social Psychology 40 (3), 385–398. [PubMed: 11593940]

Mason W, Suri S 2012 Conducting behavioral research on Amazon's Mechanical Turk. Behavioral Research Methods 44 (1), 1–23. 10.3758/s13428-011-0124-6

Matsumoto D, Ekman P Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and Neutral Faces (JACNeuF) 1988[Slides]. San Francisco, CA: Department of Psychology, San Francisco State University.

Meuwissen AS, Anderson JE, Zelazo PD, 2017 The creation and validation of the developmental Emotional Faces Stimulus Set. Behavior Research Methods 49 (3), 960–966. [PubMed: 27325165]

Minear M, Park DC, 2004 A lifespan database of adult stimuli. Behavior Research Methods, Instruments, & Computers 36 (4), 630–33. 10.3758/BF03206543

Palermo R, Coltheart M, 2004 Photographs of facial expression: accuracy, response times, and ratings of intensity. Behav Res Methods Instrum Comput 36 (4), 634–8. [PubMed: 15641409]

Pantic M, Valstar M, Rademaker R, Maat L, 2005 Web-based database for facial expression analysis, in: IEEE international Conference on Multimedia and Expo (ICME).

Paolacci G, Chandler J, 2014 Inside the Turk: Understanding Mechanical Turk as a participant pool. Current Directions in Psychological Science 23 (3), 184–8. 10.1177/0963721414531598

Phelps EA, O'Connor KJ, Cunningham WA, Funayama ES, Gatenby JC, Gore JC, Banaji MR, 2000 Performance on indirect measures of race evaluation predicts amygdala activation. Journal of Cognitive Neuroscience 12 (5), 729–738. 10.1162/089892900562552 [PubMed: 11054916]

Phelps EA, Cannistraci CJ, Cunningham WA, 2003 Intact performance on an indirect measure of race bias following amygdala damage. Neuropsychologia 41 (2), 203–208. 10.1016/S0028-3932(02)00150-1 [PubMed: 12459218]

Phillips JP, Wechsler H, Huang J, Rauss PJ, 1998, The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 16 (5), 295–306. 10.1016/S0262-8856(97)00070-X

PubMed [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005–. [Updated 2017 Aug 29]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/

RStudio Team (2016). RStudio: Integrated Development for R RStudio, Inc, Boston, MA URL http://www.rstudio.com/

Righi G, Peissig JJ, Tarr MJ, 2012 Recognizing disguised faces. Visual Cognition, 20 (2), 143–169. 10.1080/13506285.2012.654624

Rubien-Thomas E, Nardos B, Cohen AO, Li A, Cervera A, Lowery A, Dellarco DV, Rheinshmidt-Same M, Daumeyer N, Camp N, Hughes BL, Taylor-Thompson KA, Eberhardt JL, Fair DA, Richeson JA, Casey BJ, 2016 Behavioral and neural correlates of cognitive control during out-group encounters under threat. Society for Neuroscience, San Diego.

Russell JA, 1994 Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychological Bulletin 115 (2), 102–141. 10.1037/0033-2909.115.1.102 [PubMed: 8202574]

Russell JA, Bullock M, 1985 Multidimensional scaling of emotional facial expressions: similarities from preschoolers to adults. Journal of Personality and Social Psychology 48 (5), 1290–1298. 10.1037/0022-3514.48.5.1290

Samuelsson H, Jarnvik K, Henningsson H, Andersson J, Carlbring P, 2012 The Umeå university database of facial expressions: a validation study. Journal of Medical Internet Research 14 (5), e136 10.2196/jmir.2196 [PubMed: 23047935]

Strauss E, Moscovitch M, 1981 Perception of facial expressions. Brain and Language 13 (2), 308–332. 10.1016/0093-934X(81)90098-5 [PubMed: 7260576]

Tanaka JW, Kiefer M, Bukach CM, 2004 A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study. Cognition 93 (1), B1–9. 10.1016/j.cognition.2003.09.011 [PubMed: 15110726]

Tarr MJ, 2013 Face Place Retrieved November 21, 2017, from http://wiki.cnbc.cmu.edu/Face_Place

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey BJ, Nelson C, 2009 The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Research 168 (3), 242–49. 10.1016/j.psychres.2008.05.006 [PubMed: 19564050]

Tran M, Cabral L, Patel R, Cusack R, 2017 Online recruitment and testing of infants with Mechanical Turk. Journal of Experimental Child Psychology 156, 168–78. 10.1016/j.jecp.2016.12.003 [PubMed: 28088051]

Wang L, Markham R, 1999 The development of a series of photographs of Chinese facial expressions of emotion. Journal of Cross-Cultural Psychology 30 (4), 397–410. 10.1177/0022022199030004001

Warnes GR, Bolker B, Gorjanc G, Grothendieck G, Korosec A, Lumley T, MacQueen D, Magnusson A, Rogers J, and others, 2015 gdata: Various R programming tools for data manipulation. R package version 2.17.0 https://CRAN.R-project.org/package=gdata

Wickham H, 2007 Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wickham H, 2009 ggplot2: Elegant graphics for data analysis Springer-Verlag New York.

Wickham H, 2011 The split-apply-combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1–29. http://www.jstatsoft.org/v40/i01/

Winston JS, Strange BA, O'Doherty J, Dolan RJ, 2002 Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nature Neuroscience 5 (3), 277–283. 10.1038/nn816 [PubMed: 11850635]

Figure 1.

Examples of the 16 expressions in color and black and white. From top left: Angry (closed), Angry (open), Calm (closed), Calm (open), Disgust (open), Disgust (closed), Fear (closed), Fear (open), Happy (closed), Happy (open), Happy (exuberant), Neutral (closed), Neutral (open), Sad (closed), Sad (open), Surprise

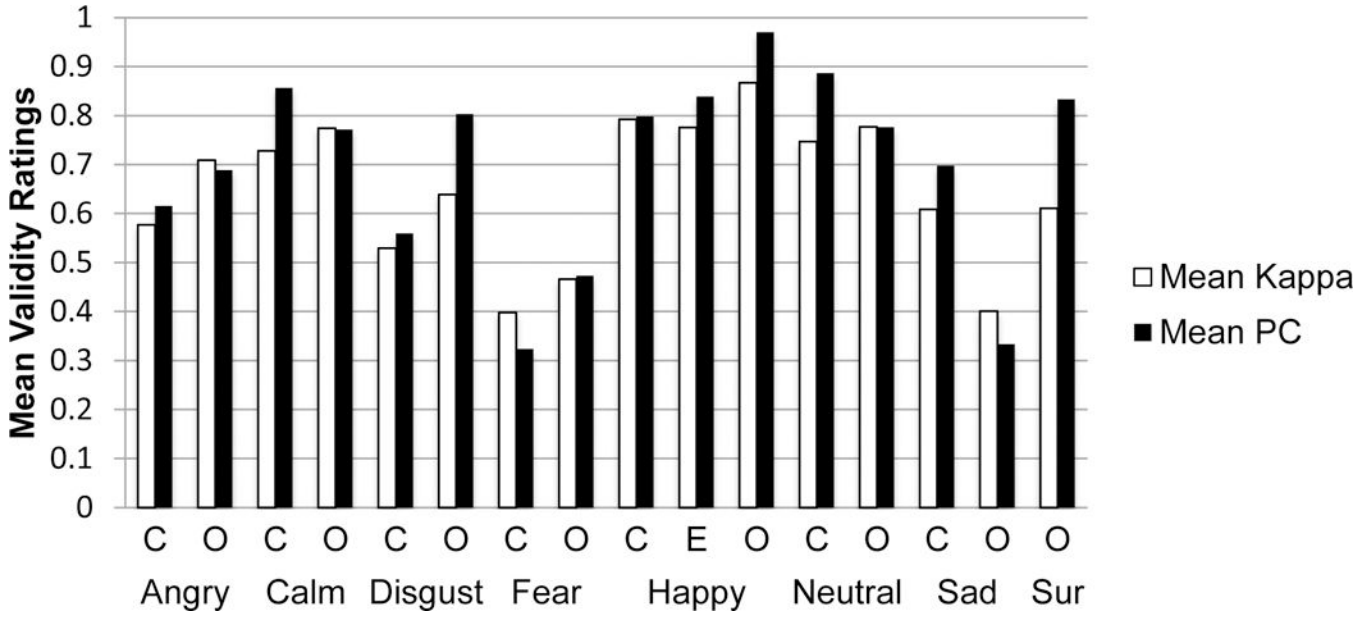

Figure 2.

Mean validity ratings for each emotional expression. C= Closed Mouth, O= Open Mouth, E= Exuberant, Sur= Surprise. PC = proportion correct.

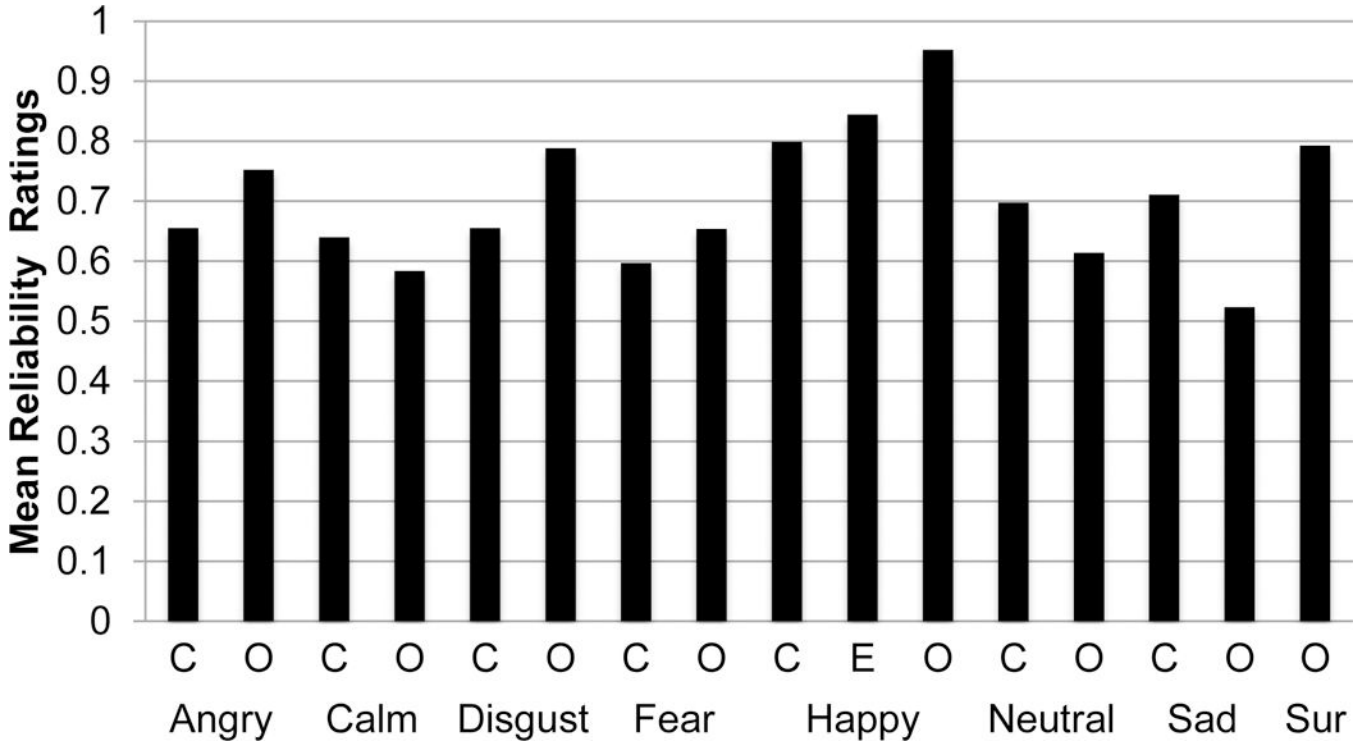

Figure 3.
Mean reliability ratings for each emotional expression. C= Closed Mouth, O= Open Mouth, E= Exuberant, Sur= Surprise

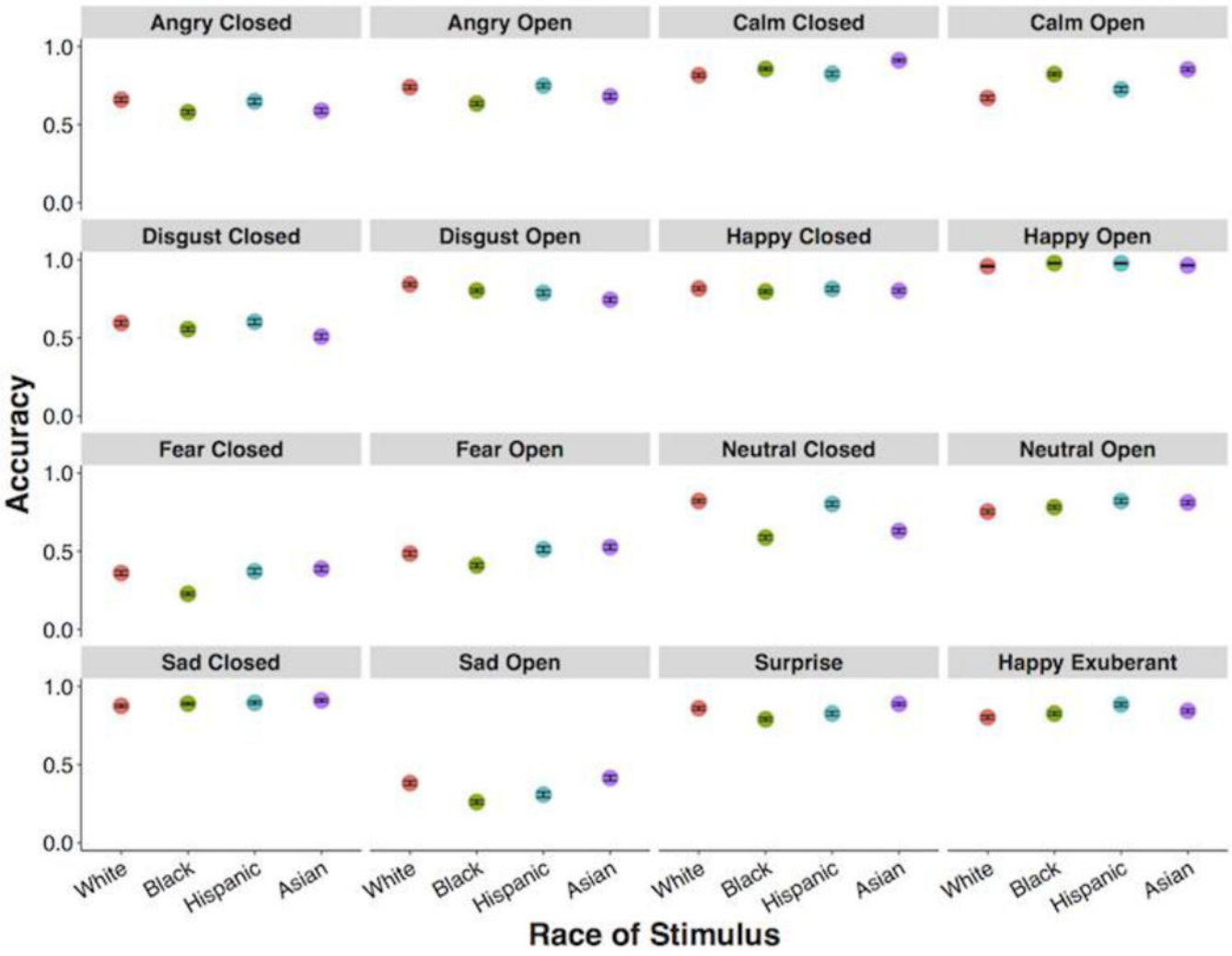

Figure 4.

Accuracy by emotion and race of stimulus across all raters (n= 651). Points represent mean accuracy with standard error bars.

Table 1

Validity of ratings for emotional expression categories.

Expression	Median proportion correct	Mean (S.D.) proportion correct	Range proportion correct	Mean (S.D.) kappa		
				Median kappa		Range kappa
Angry (closed)	0.66	0.62 (0.24)	0.00−0.98	0.61	0.58 (0.21)	−0.05−0.91
Angry (open)	0.74	0.69 (0.24)	0.00−1.00	0.78	0.71 (0.21)	−0.03−0.96
Calm (closed)	0.90	0.86 (0.14)	0.20−1.00	0.76	0.73 (0.14)	0.18−0.95
Calm (open)	0.85	0.78 (0.20)	0.14−1.00	0.81	0.78 (0.17)	0.22−0.98
Disgust (closed)	0.58	0.56 (0.23)	0.10−1.00	0.56	0.53 (0.21)	0.04−0.91
Disgust (open)	0.88	0.81 (0.19)	0.24−1.00	0.66	0.64 (0.17)	−0.06−0.90
Fear (closed)	0.30	0.33 (0.21)	0.00−0.70	0.41	0.40 (0.23)	−0.03−0.81
Fear (open)	0.51	0.48 (0.22)	0.00−0.94	0.50	0.47 (0.21)	−0.02−0.97
Happy (closed)	0.88	0.80 (0.20)	0.08−1.00	0.83	0.80 (0.17)	0.11–0.98
Happy (exuberant)	0.90	0.84 (0.20)	0.00−1.00	0.82	0.78 (0.20)	−0.01−1.00
Happy (open)	0.99	0.98 (0.06)	0.58−1.00	0.92	0.87 (0.13)	0.30−1.00
Neutral (closed)	0.94	0.89 (0.13)	0.42−1.00	0.78	0.75 (0.13)	0.38−0.95
Neutral (open)	0.85	0.78 (0.20)	0.07−0.99	0.83	0.78 (0.17)	0.11−0.96
Sad (closed)	0.79	0.70 (0.26)	0.00−0.99	0.65	0.61 (0.23)	−0.10−0.93
Sad (open)	0.29	0.34 (0.24)	0.00−0.99	0.38	0.41 (0.25)	−0.05−0.87
Surprise (open)	0.88	0.84 (0.14)	0.20−1.00	0.62	0.62 (0.15)	0.17−0.92

Table 2

Confusion matrix depicting mean proportion correct (S.D.) for MTurk participants endorsing each of the target expressions.

Photograph	Angry	Calm/Neutral	Disgust	Fear	Happy	Sad	Surprise	None of the Above
Angry (closed)	**0.62 (0.24)**	0.05 (0.07)	0.17 (0.15)	0.01 (0.02)	0.01 (0.06)	0.10 (0.15)	0.01 (0.02)	0.04 (0.05)
Angry (open)	**0.69 (0.24)**	0.01 (0.02)	0.21 (0.18)	0.03 (0.05)	0.02 (0.09)	0.01 (0.04)	0.01 (0.03)	0.02 (0.03)
Calm (closed)	0.01 (0.02)	**0.86 (0.14)**	0.01 (0.01)	0.00 (0.01)	0.10 (0.14)	0.01 (0.03)	0.01 (0.02)	0.01 (0.01)
Calm (open)	0.01 (0.02)	**0.78 (0.20)**	0.01 (0.03)	0.01 (0.03)	0.12 (0.18)	0.02 (0.03)	0.03 (0.08)	0.02 (0.03)
Disgust (closed)	0.18 (0.17)	0.02 (0.03)	**0.56 (0.23)**	0.02 (0.03)	0.01 (0.05)	0.18 (0.22)	0.01 (0.04)	0.02 (0.04)
Disgust (open)	0.04 (0.07)	0.02 (0.08)	**0.81 (0.19)**	0.03 (0.07)	0.03 (0.07)	0.01 (0.02)	0.04 (0.06)	0.03 (0.06)
Fear (closed)	0.02 (0.07)	0.05 (0.09)	0.07 (0.08)	**0.33 (0.21)**	0.01 (0.02)	0.07 (0.13)	0.43 (0.27)	0.02 (0.03)
Fear (open)	0.02 (0.07)	0.01 (0.01)	0.05 (0.08)	**0.48 (0.22)**	0.02 (0.08)	0.01 (0.02)	0.41 (0.24)	0.01 (0.02)
Happy (closed)	0.00 (0.01)	0.18 (0.19)	0.00 (0.01)	0.00 (0.01)	**0.80 (0.20)**	0.00 (0.01)	0.00 (0.01)	0.01 (0.01)
Happy (open)	0.00 (0.01)	0.01 (0.03)	0.00 (0.01)	0.00 (0.01)	**0.98 (0.06)**	0.00 (0.01)	0.01 (0.04)	0.00 (0.01)
Happy (exuberant)	0.00 (0.01)	0.02 (0.10)	0.00 (0.01)	0.00 (0.01)	**0.84 (0.20)**	0.00 (0.01)	0.13 (0.16)	0.01 (0.01)
Neutral (closed)	0.04 (0.07)	**0.89 (0.13)**	0.01 (0.03)	0.00 (0.02)	0.01 (0.02)	0.04 (0.05)	0.00 (0.01)	0.01 (0.02)
Neutral (open)	0.03 (0.05)	**0.78 (0.20)**	0.04 (0.08)	0.03 (0.04)	0.01 (0.04)	0.04 (0.04)	0.05 (0.08)	0.03 (0.05)
Sad (closed)	0.06 (0.12)	0.11 (0.15)	0.06 (0.11)	0.03 (0.06)	0.00 (0.01)	**0.70 (0.26)**	0.01 (0.02)	0.02 (0.03)
Sad (open)	0.04 (0.07)	0.08 (0.14)	0.25 (0.22)	0.16 (0.15)	0.02 (0.08)	**0.34 (0.24)**	0.07 (0.09)	0.05 (0.06)
Surprise (open)	0.00 (0.01)	0.01 (0.03)	0.01 (0.02)	0.06 (0.07)	0.07 (0.12)	0.00 (0.01)	**0.84 (0.14)**	0.01 (0.03)

Table 3

Reliability for emotional expressions between ratings 1 and 2.

Emotion	Mean Proportion correct Rating 1 (S.D)	Mean Proportion correct Rating 2 (S.D)	Agreement between Ratings 1 and 2 (S.D)
Angry (closed)	0.62 (0.24)	0.62 (0.23)	0.65 (0.15)
Angry (open)	0.69 (0.24)	0.72 (0.23)	0.75 (0.15)
Calm (closed)	0.86 (0.14)	0.87 (0.13)	0.64 (0.10)
Calm (open)	0.78 (0.20)	0.78 (0.19)	0.58 (0.11)
Disgust (closed)	0.56 (0.23)	0.55 (0.23)	0.65(0.12)
Disgust (open)	0.81 (0.19)	0.82 (0.19)	0.79 (0.15)
Fear (closed)	0.33 (0.21)	0.36 (0.22)	0.60 (0.12)
Fear (open)	0.48 (0.22)	0.52 (0.21)	0.65 (0.10)
Happy (closed)	0.80 (0.20)	0.78 (0.19)	0.80 (0.14)
Happy (exuberant)	0.84 (0.20)	0.86 (0.16)	0.84 (0.12)
Happy (open)	0.98 (0.06)	0.97 (0.05)	0.95 (0.05)
Neutral (closed)	0.89 (0.13)	0.90 (0.12)	0.70 (0.09)
Neutral (open)	0.78 (0.20)	0.81 (0.17)	0.61 (0.11)
Sad (closed)	0.70 (0.26)	0.71 (0.26)	0.71 (0.18)
Sad (open)	0.34 (0.24)	0.37 (0.37)	0.52 (0.13)
Surprise (open)	0.84 (0.14)	0.82 (0.14)	0.79 (0.12)