Psychiatry Research 168 (2009) 242–249

Contents lists available at ScienceDirect

Psychiatry Research

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

The NimStim set of facial expressions: Judgments from untrained research participants

Nim Tottenham [a,*], James W. Tanaka [b], Andrew C. Leon [c], Thomas McCarry [a], Marcella Nurse [a], Todd A. Hare [a], David J. Marcus [d], Alissa Westerlund [e], BJ Casey [a], Charles Nelson [e]

[a] Sackler Institute for Developmental Psychobiology, Weill Cornell Medical College, New York, NY, USA
[b] Department of Psychology, University of Victoria, Victoria, British Columbia, Canada
[c] Department of Psychiatry, Weill Cornell Medical College, New York, NY, USA
[d] Children's National Medical Center, Rockville, MD, USA
[e] Developmental Medicine Center, Children's Hospital Boston/Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 14 May 2007
Received in revised form 24 August 2007
Accepted 4 May 2008

Keywords:
Face
Expression
Stimulus set
Emotion
Multiracial
Validity
Reliability

ABSTRACT

A set of face stimuli called the NimStim Set of Facial Expressions is described. The goal in creating this set was to provide facial expressions that untrained individuals, characteristic of research participants, would recognize. This set is large in number, multiracial, and available to the scientific community online. The results of psychometric evaluations of these stimuli are presented. The results lend empirical support for the validity and reliability of this set of facial expressions as determined by accurate identification of expressions and high intra-participant agreement across two testing sessions, respectively.

1. Introduction

The literature on human social and emotional behavior is rich with studies of face processing. An online search in databases like PubMed for "face perception" or "face processing" to date results in over 7000 relevant papers. A similar search for "facial expression" results in over 5000 hits. Integral to these studies is the availability of valid and reliable face stimuli. This paper describes a set of face stimuli (the NimStim Set of Facial Expressions — available to the scientific community at http://www.macbrain.org/resources.htm) and reports validity and reliability for this set based on ratings from healthy, adult participants.

Facial expressions and their interpretation are a topic of interest to researchers because of the links between emotional experience and facial expression. It has been argued that expressions of facial emotion have communicative value (Darwin, 1872) with phylogenic roots (Izard, 1971). Both the production (Ekman and Friesen, 1978) and the interpretation of facial expressions (Schlosberg, 1954; Russell and Bullock, 1985) have been examined empirically. Previous research using existing sets (Ekman and Friesen, 1976) has established much of what is known about face processing. However, the parameters of existing sets may not always satisfy the objectives of the experiment (Erwin et al., 1992). For example, the number of stimuli may be too few (Winston et al., 2002), the stimulus set may not have enough racial or ethnic diversity, or there might not be an appropriate comparison or baseline expression (Phillips et al., 1998). Issues like these have motivated researchers to create their own stimuli (Hart et al., 2000; Phelps et al., 2000; Gur et al., 2002; Batty and Taylor, 2003; Phelps et al., 2003; Tanaka et al., 2004). The goal of creating the NimStim Set of Facial Expressions was to provide a set of facial expressions that address these issues.

A number of features of the set are advantageous for researchers who study face expression processing. Perhaps the most important is the racial diversity of the actors. Studies often show that the race or ethnicity of a model impacts face processing both behaviorally (Elfenbein and Ambady, 2002; Herrmann et al., 2007) and in terms of the underlying neurobiology of face processing (Hart et al., 2000; Phelps et al., 2000; Golby et al., 2001; Lieberman et al., 2005). This modulation by race or ethnicity is not identified for all populations (Beaupre and Hess, 2005) and may be driven by experience (Elfenbein and Ambady, 2003) and bias (Phelps et al., 2000; Hugenberg and Bodenhausen, 2003).

To address such findings, face sets have been developed that consist of models from various backgrounds. For example, the JACFEE

* Corresponding author. Sackler Institute for Developmental Psychobiology, Weill Cornell Medical College, 1300 York Avenue, Box #140, New York, NY 10021, USA. Tel.: +1 212 746 5830; fax: +1 212 746 5755.
E-mail address: nlt2002@med.cornell.edu (N. Tottenham).

0165-1781/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2008.05.006

set (Ekman and Matsumoto, 1993–2004) consists of Japanese and Caucasian models, Mandal's (1987) set consists of Indian models, Mandal et al.'s (2001) set consists of Japanese models, the Montreal Set of Facial Displays of Emotion (Beaupre et al., 2000) consists of French Canadian, Chinese, and sub-Saharan African models, and Wang and Markham's (1999) set consists of Chinese models. Unlike these other sets, the NimStim set provides one uniform set of Asian-American, African-American, European-American, and Latino-American actors, all photographed under identical conditions. Additionally, because these actors all live in the same metropolitan city within the U.S., the subtle differences that accompany expressions when posed by individuals from different countries are minimized (Ekman and Friesen, 1969; Matsumoto et al., 2005). The NimStim set has attributes, described below, that are not typically found in other sets that include models from non-European populations.

There are four distinguishing attributes of the set that build on previously established sets. First, the NimStim set is available in color, and it contains a large number of stimuli and a large variety of facial expressions. Whereas most sets contain roughly 100 or fewer stimuli (Mandal, 1987; Ekman and Matsumoto, 1993–2004; Wang and Markham, 1999; Mandal et al., 2001), the NimStim set contains 672, consisting of 43 professional actors, each modeling 16 different facial poses, including different examples of happy, sad, disgusted, fearful, angry, surprised, neutral, and calm. Secondly, a neutral expression is included in this set. The neutral expression is sometimes included in other facial expression sets (Ekman and Friesen, 1976; Erwin et al., 1992; Ekman and Matsumoto, 1993–2004), but is often omitted, particularly in sets that include models from different racial and ethnic backgrounds (Mandal, 1987; Wang and Markham, 1999; Beaupre et al., 2000; Mandal et al., 2001). The inclusion of the neutral expression is important since neutral is often a comparison condition, particularly in neuroimaging studies (Breiter et al., 1996; Thomas et al., 2001). Thirdly, the NimStim set contains open- and closed-mouth versions of each expression, which can be useful to experimentally control for perceptual differences (e.g., toothiness) that can vary from one expression to another, as this featural difference may bias responses (Kestenbaum and Nelson, 1990). Having closed- and open-mouth versions also facilitates morphing between various expressions by reducing blending artifacts. Lastly, since positively valenced faces are generally lower in arousal than negatively valenced faces and can present an arousal/valence confound, this set includes three degrees of happy faces (e.g., closed-mouth, open-mouth, and high arousal/exuberant).

A final distinguishing feature of this set is the inclusion of a calm face. Studies examining the perception of facial expressions often use neutral faces as the comparison face (Breiter et al., 1996; Vuilleumier et al., 2001). However, there is evidence to suggest that neutral faces may not always be perceived as emotionally neutral (Donegan et al., 2003; Somerville et al., 2004; Iidaka et al., 2005), especially for children (Thomas et al., 2001; Lobaugh et al., 2006). Researchers have artificially generated other comparison faces (i.e., 25% happy) to address this concern (Phillips et al., 1997). Within the NimStim Set, a calm expression category is provided, which is perceptually similar to neutral, but may be perceived as having a less negative valence. Here we provide data validating the use of the calm face.

Validation of the entire set was accomplished by asking participants to label each stimulus. A different method of rating face stimuli involves having highly trained raters use facial action units to make judgments about the expressions (Ekman and Friesen, 1977; Ekman and Friesen, 1978). This method is very useful for establishing the uniformity of expression exemplars. The merit of the method used in this article is that the ratings were obtained from untrained volunteers, who are characteristic of those in face expression processing studies. In other words, the approach taken in this study to establish whether a certain expression was perceived as the intended expression was to measure concordance between the subjects' labels and the intended expressions posed by the actors (the validity measure) as well as the intra-participant test–retest reliability. We hypothesized that the participants' judgments of these stimuli would provide empirical support for the reliability and validity of this new set of facial expressions.

2. Method

2.1. Participants

Data were collected from two groups of participants, producing a total *N* of 81 participants. The first group included 47 undergraduate students from a liberal arts college located in the Midwestern United States. Mean age was 19.4 years (18–22, S.D. = 1.2), and the majority (39/47) of these respondents were female. The participants received course credit for their participation in the study. Based on the participant pool data, it was estimated that 81% (38/47) were European-American, 6% (3/47) were African-American, 9% (4/47) were Asian-American, and 4% (2/47) were Hispanic-American. Only validity ratings, not reliability ratings, were obtained from this first group of participants. In order to increase the number of participants for the validity ratings, in particular, male participants, we recruited a second group of participants, which comprised 34 volunteers from the New York Metropolitan area. Mean age of this group was 25.8 years (19–35, S.D. = 4.1), and 35% (12/34) were female and 65% (22/34) were male. The participants were monetarily compensated for their time. Fifty-nine percent (20/34) were European-American, 18% (6/34) were African-American, 6% (2/34) were Asian-American, 6% (2/34) were Latino-American, and 12% (4/34) identified as a non-listed race or ethnicity. Three measures were obtained from this second group of participants: validity, test–retest reliability, and calm vs. neutral validity.

2.2. Stimuli

Stimuli were images from the NimStim Set of Facial Expressions (672 images; http://www.macbrain.org/resources.htm), which consisted of naturally posed photographs (e.g., with hair, make-up) of 43 professional actors (18 female, 25 male; 21 years old–30 years old) in New York City. Actors were African- ($N = 10$), Asian- ($N = 6$), European- ($N = 25$), and Latino-American ($N = 2$). Actors were instructed to pose eight expressions: happy, sad, angry, fearful, surprised, disgusted, neutral, and calm (see Fig. 1). For each expression, separate open- and closed-mouth versions were posed, except for surprise, which were only posed with an open mouth[1]. Negatively valenced faces typically differ from faces like happy in terms of valence, but also are higher in arousal level. Therefore, three versions of happy were obtained (closed-mouth, open-mouth, and high arousal open-mouth/exuberant). All stimuli were included in this validation paradigm regardless of the quality of acting.

Actors were instructed to pose a particular expression (e.g., "Make a happy face") and produce the facial expression as they saw fit (Mandal, 1987; Mandal et al., 2001). Once one version of the facial expression (e.g., closed mouth) was created and photographed, the other version (e.g., open mouth) was prompted and photographed. To create the calm faces, actors were instructed to transfigure their neutral face into a more relaxed one, as if they were engaged in a calming activity or otherwise pleasantly preoccupied. Therefore, the calm faces were essentially neutral faces with less overall muscle tension in the face. Actors were paid for their time.

2.3. Evaluation procedure

2.3.1. Validity

Participants were seated approximately 53 cm from the computer. All 672 stimuli were presented on a Macintosh computer using the Psyscope experimental software package (Cohen et al., 1993). Images

[1] A closed-mouth version of surprise is modeled by two of the actors.

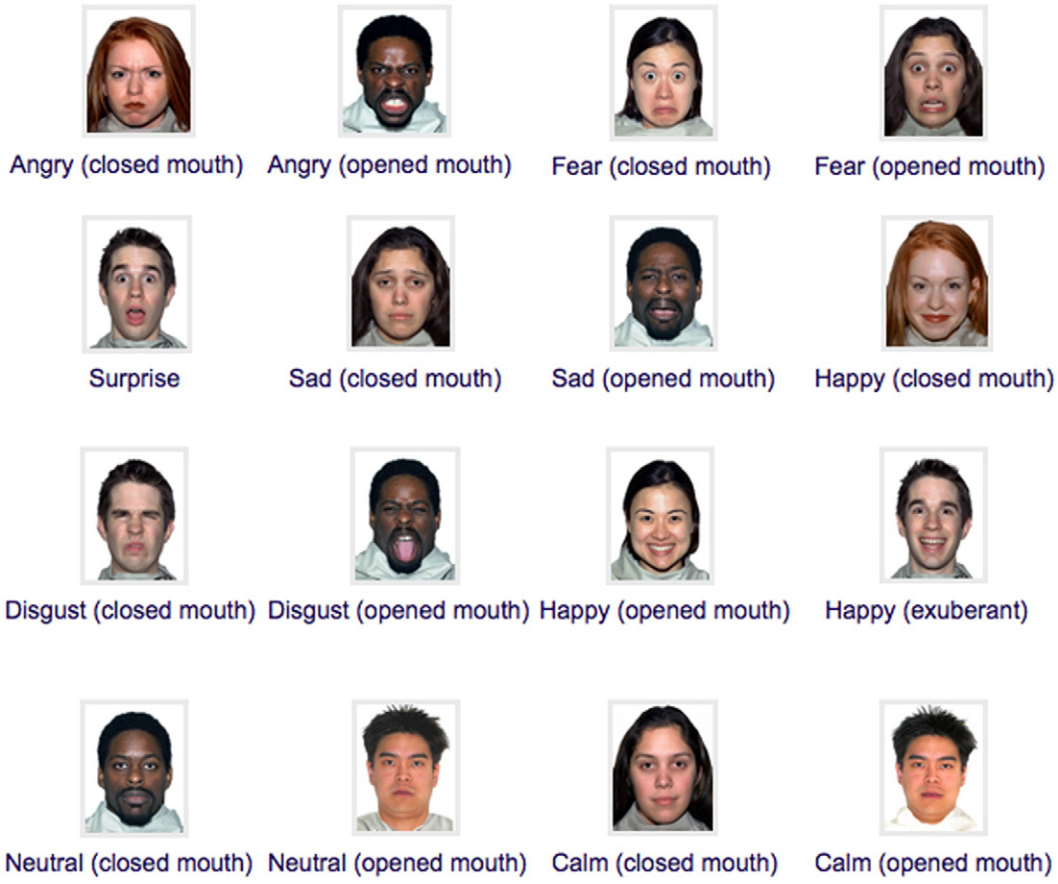

Fig. 1. Examples of the 16 expressions posed by actors.

were presented in 256-bit color at a vertical visual angle of 9.11° and a horizontal angle of 7°. On each trial, a face was presented with the choices "angry", "surprise", "afraid", "sad", "happy", "disgust", "neutral", "calm", and "none of the above". The participant's task was to label each image or indicate that none of the labels were appropriate (i.e., "none of the above"). Stimuli were presented randomly in a different order for each participant, and participants proceeded at their own pace.

2.3.2. Reliability

After an approximately 20-minute break following the first presentation of the stimuli, participants rated all of the 672 face stimuli for a second time using the same procedure described above for the validity ratings. Stimuli were presented in a different random order than the first presentation. Judgments made during this second presentation were not included in the validity scores.

2.3.3. Calm vs. neutral validity ratings

During the break between ratings of the set for the first and second time, participants rated calm and neutral faces, which were presented side by side, with one model posing each face per slide. The images were presented at a visual vertical angle of 14.8° and horizontal angle of 10.7° and presented grayscale[2]. These calm and neutral faces were the same ones included in the judgments of the entire set (i.e., when validity and reliability ratings were obtained). Participants were provided with these definitions and asked to label each face as calm, neutral, or neither:

> NEUTRAL: plain, alert face, like passport photo. Neither negative nor positive.
> CALM: similar to neutral, almost bordering on pleased, or slightly happy — maybe daydreaming. Person looks very serene, less threatening than neutral.

2.4. Data analytic procedures

2.4.1. Validity

Validity ratings were collected from the two groups of participants, and these ratings did not differ from each other for both proportion correct and kappa scores (see Supplemental Table 1a and 1b); therefore, their data were combined[3]. These validity scores were calculated from the first presentation of the 672 stimuli. Evaluations of facial expression sets typically rely on percent correct (i.e., proportion of participants who agree with intended expression) as the dependent variable (Ekman and Friesen, 1976; Mandal, 1987; Biehl et al., 1997; Wang and Markham, 1999; Mandal et al., 2001; Beaupre and Hess, 2005). However, this type of statistic does not account for false positives (Erwin et al., 1992). Although proportion correct is reported here, we also report Cohen's kappa (Cohen, 1960), which is a chance-corrected measure of agreement between the intended expression and the participants' labels. There were 672 stimuli resulting in 672 kappas.

[2] Although calm and neutral face stimuli are available in color, calm and neutral validity ratings were obtained using grayscale stimuli because validity of these two stimulus types needed to be established as part of a different protocol that used grayscale images.

[3] The first group of participants did not rate stimuli from model #45 whose photographs were obtained after the first group was recruited.

These kappas were calculated to estimate concordance between the labels and the intended expression for each stimulus, examined within actor, separately for open- and closed-mouth conditions[4]. Since the 'none of the above" option was provided, this choice was considered incorrect in the kappa calculation just as other incorrect emotion labels were. Lastly, it was not anticipated that calm and neutral would be identified as two separate expressions during this labeling task since the perceptual differences between the faces are quite small relative to the differences between the other expressions tested. Therefore, responses of "calm" or "neutral" were accepted as correct for both calm and neutral facial expressions.

2.4.2. Reliability

Because of the high test–retest reliability for most of the stimuli, calculating kappa was often mathematically intractable and did not appropriately reflect the concordance of ratings over time[5]. Therefore, proportion agreement between the first and second ratings was used to quantify extent of reliability of each face stimulus.

2.4.3. Validity of calm and neutral

As with the validity ratings of the entire data set, we calculated proportion correct and kappas for each calm and neutral stimulus to estimate the concordance between emotion label (either "calm", "neutral", or "neither") and intended expression within each actor. The answer "neither" was also counted as an incorrect response.

3. Results

3.1. Validity

There were two validity measures (proportion correct and kappa scores) for each stimulus, thus resulting in 672 proportion correct and kappa scores. These 672 proportion correct and kappa scores are presented individually in Supplemental Tables 2a and 2b and in aggregate by actor (43 actors × 2 mouth positions = 86 scores) in Supplemental Table 3, but by expression (17 scores) for succinctness within this manuscript (Table 1; closed and open mouth separately). The overall proportion correct was high (mean = 0.81 (S.D. = 0.19), median = 0.88). The overall concordance between raters' labels and the intended expressions was also high (mean kappa across stimuli = 0.79 (S.D. = 0.17); median kappa = 0.83). Table 2 presents the confusion matrix for the labels chosen by participants, which represents the average proportion of target and non-target labels endorsed for each expression and shows that errors were fairly consistent for each expression (e.g., incorrect judgments for fear faces were usually mislabeled as "surprise").

Validity ratings were similar from one actor to another. Forty-five percent (39/86) of the kappas calculated per actor ranged from 0.8 to 1.0, a range considered to reflect almost perfect concordance between the given label and intended expression by Landis and Koch (1977), and the corresponding mean proportion correct scores ranged from 0.81 to 0.93. The remaining 55% (47/86) of these kappas ranged from 0.59 and 0.79, and the mean proportion correct scores ranged from 0.67 to 0.84.

Table 1
Description of validity ratings for individual emotional expressions ($N = 81$ subjects rating 672 stimuli).

	Median proportion correct	Mean (S.D.) proportion correct	Median kappa	Mean (S.D.) kappa
Angry (closed)	0.90	0.84 (0.17)	0.81	0.78 (0.13)
Calm (closed)	0.90	0.88 (0.07)	0.87	0.84 (0.09)
Disgust (closed)	0.86	0.76 (0.23)	0.83	0.75 (0.19)
Fear (closed)	0.51	0.47 (0.21)	0.58	0.54 (0.20)
Happy (closed)	0.94	0.92 (0.07)	0.94	0.92 (0.06)
Neutral (closed)	0.93	0.91 (0.06)	0.87	0.86 (0.08)
Sad (closed)	0.91	0.83 (0.16)	0.76	0.76 (0.13)
Surprised (closed)	0.61	0.61 (0.10)	0.62	0.62 (0.08)
Angry (open)	0.96	0.90 (0.15)	0.92	0.88 (0.11)
Calm (open)	0.81	0.79 (0.11)	0.84	0.81 (0.10)
Disgust (open)	0.93	0.84 (0.21)	0.82	0.77 (0.18)
Fear (open)	0.74	0.73 (0.12)	0.69	0.67 (0.12)
Happy (open)	0.99	0.98 (0.02)	0.97	0.95 (0.05)
Neutral (open)	0.86	0.82 (0.11)	0.86	0.83 (0.09)
Sad (open)	0.59	0.60 (0.21)	0.64	0.62 (0.18)
Surprised (open)	0.86	0.81 (0.13)	0.68	0.68 (0.12)
Happy (open exuberant)	0.88	0.86 (0.13)	0.90	0.88 (0.09)

There were differences in proportion correct ($F(15, 435) = 24.65$, $P < 0.0001$) and kappa scores ($F(15, 435) = 28.65$, $P < 0.0001$) from one emotional expression to another indicating that some emotions were more accurately identified than others. Approximately half (8/17) of the mean kappas for emotional expression were above a 0.8 kappa cut point. The expressions above this threshold included happy (open), happy (closed), angry (open), happy (open exuberant), neutral (closed), calm (closed), neutral (open), and calm (open), and these expressions had mean proportion correct scores that ranged between 0.79 and 0.98. Another 47% (8/17) of the expressions had mean kappas between 0.6 and 0.79. These expressions included angry (closed), disgust (open), sad (closed), disgust (closed), surprised (open), fear (open), sad (open), and surprised (closed), and these expressions had mean proportion correct scores that ranged from 0.61 to 0.84. The fear (closed) expression had a mean kappa of 0.54 and a mean proportion correct score of 0.47. Table 3 shows the descriptive and inferential statistics that compare the open- and closed-mouth versions of expressions. These calculations show that expressions were not identified equally for open and closed versions. Angry, fear, and happy faces resulted in higher kappa scores with open mouths, whereas sad was more accurately identified with a closed mouth.

3.2. Reliability

Reliability scores (i.e., proportion agreement) were calculated for each stimulus to quantify agreement between times 1 and 2 for each stimulus, and these values for individual stimuli can be found in Supplemental Table 4. To show the data in a succinct fashion, results are presented in aggregate for each expression (Table 4; closed and open mouth separately). Overall, there was agreement between times 1 and 2, with a mean (S.D.) reliability score of 0.84 (0.08) and median of 0.86.

There was little variability in reliability scores from one actor to another (closed mouth and open mouth calculated separately). Ninety-one percent (78/86) of the actors had mean reliability scores that ranged between 0.80 and 1.00. The remaining 9% was between 0.73 and 0.79. In contrast, the mean reliability scores for each emotional expression indicate that some emotions were more reliably identified than others (see Table 4). The majority (13/17) of expressions had

[4] These calculations involved separate 2 × 2 contingency tables for each intended emotion in which the accuracy of the ratings was cross-classified with intended emotion, examined at the level of actor. For instance, a stimulus was either *fear* or a non-fear face (e.g., happy, sad, surprise). Likewise, the corresponding label of that stimulus provided by the raters was either "fear" or not. All ratings (approximately 567 for the closed-mouth expressions (81 raters × 7 expressions) and 729 for the open-mouth expressions (81 raters × 9 expressions)) were used in each validity contingency table.

[5] The reliability paradigm resulted in quite limited variability in the cells of the contingency table (i.e., as a group, subjects rated faces nearly identically in the two sessions, resulting in very high agreement). Kappa is not appropriate for certain cases such as when there is very low (or very high) prevalence of events (Kraemer et al., 2002. Tutorial in biostatistics: Kappa coefficients in medical research. Statistics in Medicine 21, 2109–2129).

Table 2
Confusion matrix for mean (S.D.) proportion of subjects who endorsed each emotion label.

	Label								
Photograph	Angry	Calm/neutral	Disgust	Fear	Happy	Sad	Surprised	Nothing	*N*
Angry (closed)	**0.84 (0.17)**	0.03 (0.07)	0.05 (0.09)	0.01 (0.02)	0.00 (0.01)	0.05 (0.08)	0.00 (0.01)	0.01 (0.02)	78.93 (7.22)
Angry (open)	**0.90 (0.15)**	0.00 (0.01)	0.05 (0.09)	0.02 (0.05)	0.00 (0.01)	0.01 (0.02)	0.00 (0.01)	0.01 (0.02)	79.07 (7.27)
Calm (closed)	0.02 (0.03)	**0.88 (0.07)**	0.00 (0.01)	0.01 (0.01)	0.02 (0.05)	0.02 (0.05)	0.00 (0.01)	0.02 (0.02)	79.33 (7.36)
Calm (open)	0.01 (0.02)	**0.79 (0.11)**	0.02 (0.02)	0.03 (0.05)	0.01 (0.04)	0.03 (0.05)	0.05 (0.05)	0.05 (0.04)	78.95 (7.31)
Disgust (closed)	0.13 (0.18)	0.01 (0.01)	**0.76 (0.23)**	0.01 (0.01)	0.00 (0.01)	0.08 (0.17)	0.00 (0.01)	0.01 (0.02)	79.15 (7.47)
Disgust (open)	0.03 (0.04)	0.00 (0.01)	**0.84 (0.21)**	0.02 (0.03)	0.00 (0.01)	0.09 (0.20)	0.01 (0.02)	0.01 (0.02)	79.14 (7.11)
Fear (closed)	0.03 (0.06)	0.04 (0.06)	0.04 (0.08)	**0.47 (0.21)**	0.01 (0.01)	0.10 (0.16)	0.29 (0.20)	0.02 (0.02)	78.92 (7.76)
Fear (open)	0.02 (0.05)	0.00 (0.01)	0.02 (0.03)	**0.73 (0.12)**	0.01 (0.01)	0.01 (0.03)	0.19 (0.12)	0.01 (0.01)	79.37 (7.13)
Happy (closed)	0.00 (0.01)	0.06 (0.06)	0.00 (0.01)	0.00 (0.00)	**0.92 (0.07)**	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	79.31 (7.21)
Happy (open)	0.00 (0.01)	0.01 (0.02)	0.00 (0.00)	0.00 (0.00)	**0.98 (0.02)**	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	79.47 (7.13)
Happy (open exuberant)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	**0.86 (0.13)**	0.00 (0.00)	0.13 (0.12)	0.00 (0.01)	79.31 (7.20)
Neutral (closed)	0.02 (0.03)	**0.91 (0.06)**	0.00 (0.01)	0.01 (0.01)	0.01 (0.01)	0.03 (0.05)	0.00 (0.01)	0.02 (0.02)	79.35 (7.12)
Neutral (open)	0.02 (0.02)	**0.82 (0.11)**	0.01 (0.02)	0.04 (0.05)	0.01 (0.01)	0.03 (0.05)	0.05 (0.05)	0.03 (0.02)	79.14 (7.34)
Sad (closed)	0.03 (0.05)	0.04 (0.06)	0.04 (0.07)	0.03 (0.07)	0.00 (0.01)	**0.83 (0.16)**	0.01 (0.04)	0.02 (0.02)	79.05 (7.17)
Sad (open)	0.02 (0.04)	0.07 (0.13)	0.15 (0.14)	0.09 (0.10)	0.00 (0.01)	**0.60 (0.21)**	0.03 (0.03)	0.03 (0.03)	79.10 (7.33)
Surprised (closed)	0.00 (0.00)	0.12 (0.05)	0.00 (0.00)	0.18 (0.06)	0.06 (0.08)	0.01 (0.01)	**0.61 (0.10)**	0.02 (0.02)	80.50 (0.71)
Surprised (open)	0.00 (0.00)	0.01 (0.00)	0.00 (0.01)	0.14 (0.13)	0.03 (0.07)	0.00 (0.00)	**0.81 (0.13)**	0.00 (0.01)	79.35 (7.11)

Target emotion in bold.

average reliability scores that ranged between 0.80 and 1.00, and the remaining 24% ranged between 0.68 and 0.77.

3.3. Calm and neutral validity

Eighty-six proportion correct scores and kappas were calculated for the calm and neutral stimuli, one for each calm or neutral stimulus (see Table 5). Scores varied considerably from one actor to another indicating that some actors expressed calm and neutral faces better than others. Mean proportion correct for neutral faces was 0.72 (S.D. = 0.18; range: 0.12–0.91), median proportion correct was 0.76, mean kappa was 0.34 (S.D. = 0.39; range: −0.65–0.88), and median kappa was 0.38. Mean proportion correct for calm faces was 0.56 (S.D. = 0.27; range: 0.09–0.91), median proportion correct was 0.62, mean kappa was 0.32 (S.D. = 0.40; range: −0.65–0.88), and median kappa was 0.38. Nearly half of the neutral faces (21/43) and half of the calm faces (21/43) had kappas that exceeded 0.40, and these faces had a mean proportion correct of 0.82 and 0.79, respectively. The proportion correct scores for calm and neutral faces are positively correlated ($r(41) = 0.69$, $P<0.001$), indicating that a model who posed an identifiable neutral face was also likely to pose an identifiable calm face.

4. Discussion

The purpose of this article was to present a set of facial expression stimuli and to present data describing the judgments of these stimuli by untrained, healthy adult research participants. This face set is large in number, is multiracial, has contemporary looking, professional actors, contains a variety of expressions, and is available to the scientific community online. For these reasons, this set may be a resource for scientists who study face perception.

Validity was indexed as how accurate participants were at identifying each emotional expression, and these scores were high. We examined proportion correct, as others have, in order to compare our results with those from other sets. The calculations included ratings from the entire set, and therefore included stimuli of both high and low posing quality. Nonetheless, the overall mean proportion correct obtained with this set was 0.79, which is well above the 0.70 criterion of the other sets that include models from non-European backgrounds. The scores are comparable to those reported for the Pictures of Facial Affect (Ekman and Friesen, 1976), where the mean accuracy was 88%, and for the JACFEE set (Ekman and Matsumoto, 1993–2004), where the average percent correct was 74% (Biehl et al., 1997).

Different face sets have provided different response options to participants. As Russell (1994) points out, the response option can bias the level of accuracy obtained in expression recognition studies. Forced choice methods, like those commonly used in validating face

Table 3
Comparing closed and opened mouth versions of five basic emotional expressions.

	Closed mean	Open mean	*t*	Df	Sig. (2-tailed)
Kappa					
Angry	0.78	0.88	−4.60	42	0.00
Disgust	0.75	0.76	−0.25	38	0.80
Fear	0.54	0.67	−4.12	35	0.00
Happy	0.92	0.95	−3.13	41	0.00
Sad	0.76	0.63	4.25	40	0.00
Proportion correct					
Angry	0.84	0.90	−1.90	42	0.06
Disgust	0.76	0.84	−1.73	38	0.09
Fear	0.47	0.73	−7.47	35	0.00
Happy	0.92	0.98	−5.08	41	0.00
Sad	0.83	0.60	6.85	40	0.00

Table 4
Description of reliability for individual emotional expressions ($N = 34$ subjects, from group 2 only — rating 672 stimuli).

Emotion	Proportion correct Block 1	Proportion correct Block 2	Agreement between Blocks 1 and 2
Angry (closed)	0.86	0.88	0.87
Calm (closed)	0.89	0.96	0.90
Disgust (closed)	0.74	0.75	0.81
Fear (closed)	0.46	0.52	0.68
Happy (closed)	0.93	0.92	0.91
Neutral (closed)	0.92	0.99	0.94
Sad (closed)	0.84	0.84	0.85
Surprised (closed)	0.59	0.58	0.73
Angry (open)	0.91	0.92	0.90
Calm (open)	0.81	0.91	0.85
Disgust (open)	0.84	0.84	0.87
Fear (open)	0.75	0.79	0.75
Happy (open)	0.98	0.99	0.98
Neutral (open)	0.84	0.93	0.86
Sad (open)	0.62	0.66	0.77
Surprised (open)	0.83	0.84	0.80
Exuberant happy	0.88	0.86	0.87

Note — there is no model #4 or 44.

Table 5
Description of validity ratings for semi-forced choice calm vs. neutral discrimination ($N = 34$ subjects rating 86 stimuli).

Model	Calm		Neutral	
	Proportion correct	Kappa	Proportion correct	Kappa
1	0.62	0.50	0.88	0.56
2	0.44	0.12	0.65	0.15
3	0.94	0.82	0.88	0.82
5	0.18	−0.12	0.62	−0.12
6	0.71	0.47	0.74	0.50
7	0.71	0.56	0.85	0.59
8	0.68	0.62	0.76	0.50
9	0.12	−0.56	0.26	−0.56
10	0.79	0.59	0.76	0.56
11	0.88	0.68	0.76	0.65
12	0.76	0.68	0.85	0.62
13	0.41	0.26	0.74	0.26
14	0.35	0.18	0.79	0.29
15	0.47	0.21	0.71	0.26
16	0.88	0.79	0.91	0.79
17	0.68	0.38	0.68	0.38
18	0.85	0.65	0.79	0.68
19	0.38	0.18	0.76	0.32
20	0.47	0.26	0.76	0.29
21	0.44	−0.09	0.44	−0.09
22	0.94	0.85	0.91	0.88
23	0.91	0.82	0.88	0.82
24	0.91	0.74	0.79	0.74
25	0.62	0.44	0.76	0.41
26	0.26	0.06	0.76	0.12
27	0.68	0.56	0.85	0.62
28	0.85	0.71	0.82	0.68
29	0.26	0.09	0.76	0.24
30	0.24	0.12	0.76	0.09
31	0.35	0.03	0.65	0.03
32	0.94	0.88	0.88	0.82
33	0.12	−0.62	0.26	−0.59
34	0.21	−0.29	0.47	−0.32
35	0.32	0.15	0.79	0.35
36	0.38	0.18	0.74	0.21
37	0.56	0.41	0.71	0.44
38	0.09	−0.24	0.59	−0.26
39	0.79	0.62	0.79	0.62
40	0.62	0.35	0.71	0.32
41	0.79	0.53	0.74	0.56
42	0.44	0.03	0.56	0.09
43	0.18	−0.65	0.12	−0.65
45	0.85	0.76	0.88	0.74

Note — there is no model #4 or 44.

expressions sets, can inflate accuracy because these procedures bias the participant towards a particular hypothesis. However, the other extreme of freely chosen labels is also not ideal because participants tend to provide scenarios (e.g., "she saw a ghost") rather than expressions (e.g., fear) or, as a group, they rarely choose the same word to describe the expression, forcing researchers to make judgments regarding individual responses. The current study chose a semi-forced choice method that was less strict than forced choice while being more interpretable than the free label method. Participants were provided with a "none of the above" option. As a result, participants labeled only those faces that they felt met a criterion of adequately conveying an emotional expression. In light of this more stringent procedure, the high accuracy scores obtained from the NimStim set are all the more striking.

Calculating proportion correct does not account for false alarms (e.g., the number of times participants called a non-fear face "fear"), and therefore, we also calculated kappa for each stimulus to measure concordance between participants' labels and the intended expressions. Landis and Koch (1977) have defined "moderate" concordance as kappas with values between 0.4 and 0.6, "substantial" concordance between 0.6 and 0.8, and "almost perfect" concordance of 0.8 and greater. The mean kappa obtained in this set was 0.79, which was well within the "substantial" range. Kappas calculated for each actor were all in the "substantial" and "almost perfect" range. Only one actor resulted in an average kappa below the "substantial" range; that actor's images have been removed from the set, and now Version 2.0 of the NimStim set is available online. This set only contains stimuli from actors with kappas within the "substantial" and "almost perfect" range.

The high variability in scores across emotion categories was expected since emotion recognition differs across expressions (Strauss and Moscovitch, 1981; Calder et al., 2003), and this variability has been shown in other sets (Ekman and Friesen, 1976). It is unlikely that the source of the variability is the posed nature of the images since the same variability is observed when viewers judge spontaneously produced faces (Gur et al., 2002). Typically and here, happy expressions have high recognition rates, and negative expressions (in particular, sad, fear, and surprised) have poor recognition rates. There are many hypotheses one could generate regarding the variability. The greater accuracy in recognizing happy faces may be a result of greater familiarity with happy faces or the result of the rewarding aspects of happy faces (Hare et al., 2005). While this article cannot address the cause of inter-expression variability, the results from this study replicate what has been shown in other studies where happy expressions are accurately recognized and negative expressions are poorly recognized (Biehl et al., 1997; Lenti et al., 1999; Gur et al., 2002; Elfenbein and Ambady, 2003).

Reliability was indexed by comparing judgments from time 1 to judgments from time 2. Kappas were not the appropriate statistic for these ratings since agreement was very high and the corresponding rates of disagreement were too low (Kraemer et al., 2002); therefore, we calculated proportion agreement between times 1 and 2 for each stimulus. Test–retest reliability was high as measured by proportion agreement. There was variability from one expression to another and little variability from one actor to another. No other study has reported test-retest reliability for judgments of face stimuli, so it is not possible to say whether this pattern of reliability is common. These data suggest that the stimuli are rated consistently across multiple sessions.

The method of expression creation itself can bias interpretation. Models could be instructed to move certain muscle groups to produce an expression (Ekman and Friesen, 1976), which produces uniform expressions across models, but might jeopardize the ecological validity of the images (Russell, 1994). On the other hand, naturally occurring facial expressions might lead to greater authenticity in the images, but these types of images can result in greater variability from one stimulus to another (Russell, 1994), which may not be ideal for experimental paradigms. To maintain the natural variability across models while maintaining some degree of uniformity across exemplars so viewers easily interpret them, the actors in the NimStim set were given an emotion category and instructed to create the expression themselves (Mandal, 1987). The results of this study demonstrate that this method, like other posed methods (Ekman and Friesen, 1976), results in highly accurate judgments, although the ecological validity of these faces cannot be determined by the current study. There is growing interest in the dynamics of facial expressions (Ambadar et al., 2005), and the stimuli presented in this article are deliberate, strong in intensity, and displayed as static photographs. While they may at times exaggerate aspects of naturally occurring facial expressions, the merit of self-posed expressions is that they make possible the creation of a large bank of uniform stimuli.

Because of concerns regarding the emotional "neutrality" of neutral faces (Thomas et al., 2001; Donegan et al., 2003), this set of facial expressions included a calm face. The intention was to create a facial expression that participants would explicitly label as neutral/plain, but may actually interpret as a less emotionally significant face

relative to a neutral face. Scores were lower for the calm/neutral validity ratings relative to ratings of the entire set, which was expected considering the increased difficulty in judging the difference between calm and neutral. Despite this level of difficulty, a significant proportion of the neutral and calm faces were correctly labeled above chance levels (although neutral was correctly identified more often than calm). These expressions were posed consistently within actor such that when an actor posed a calm face well, he or she also posed a neutral face well, making a complete set of calm and neutral faces available to researchers.

There are some shortcomings of the set. First, although the rating system used here was conservative (i.e., not a forced choice paradigm), it would be even more informative if future studies employed a system of responding where participants rated faces on a continuum for each type of emotion label. This "quantitative rating" (Russell, 1994) approach is a sensitive measure to capture subtle combinations of emotions within a face. Given the large number of images in the set, collecting data in this quantitative manner would preclude a within-participants design (e.g., instead, group A would rate the first third, group B would rate the second third, and group C would rate the third), and a between-participants design would weaken conclusions made regarding the consensus of the interpretations. A second issue concerns the open-/closed-mouth versions, which were created to control for the strong perceptual feature of an open mouth. We tested recognition across both open- and closed-mouth versions of each expression, and on average, this manipulation disrupts recognition of some facial expressions. However, on an individual model level, there are expressions that are accurately identified with both open- and closed-mouth versions. So, on average manipulating the mouth while maintaining the intended expression is difficult for most models to do, but there are individual actors who produce both versions well. Thirdly, unlike other stimulus sets that strip faces of extra paraphernalia (Erwin et al., 1992), actors were not instructed to remove make-up, jewelry, or facial hair. This decision could bias judgments, but it also results in a set of faces that are more representative of faces people see every day. Finally, while images from this set are available for the scientific community for use in experiments at no cost, only a subset the images may be printed in scientific publications, and these models are listed online at http://www.macbrain.org/resources.htm. The remaining models may not be published in any form.

The goal in creating this set of facial expressions was to provide a large, multiracial set of photos of professional actors, posing expressions that untrained experimental participants could identify. Having untrained participants identify the emotions on the faces is the best way to assess how similar participants in future studies will interpret the expressions. The data presented in this article should raise experimenters' confidence level about both the validity and reliability of these expressions in so far as untrained participants in a face processing study perceive them. The versatility of this set makes it a useful resource to ask new questions and draw conclusions about social perception that are generalizable to a broader range of faces.

Acknowledgements

The research reported was supported by the John D. & Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development (C.A. Nelson, Chair).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2008.05.006.

References

Ambadar, Z., Schooler, J.W., Cohn, J.F., 2005. Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. Psychological Science 16, 403–410.

Batty, M., Taylor, M.J., 2003. Early processing of the six basic facial emotional expressions. Cognitive Brain Research 17, 613–620.

Beaupre, M.G., Hess, U., 2005. Cross-cultural emotion recognition among Canadian ethnic groups. Journal of Cross-Cultural Psychology 36, 355–370.

Beaupre, M.G., Cheung, N., Hess, U., 2000. The Montreal Set of Facial Displays of Emotion. Montreal, Quebec, Canada.

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, 3–21.

Breiter, H., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S., Buckner, R.L., Strauss, M.M., Hyman, S.E., Rosen, B.R., 1996. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17, 875–887.

Calder, A.J., Keane, J., Manly, T., Sprengelmeyer, R., Scott, S., Nimmo-Smith, I., Young, A. W., 2003. Facial expression recognition across the adult life span. Neuropsychologia 41, 195–202.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Cohen, J.D., MacWhinney, B., Flatt, M., Provost, J., 1993. PsyScope: a new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments, and Computers. Behavioral Research Methods, Instruments, and Computers 25, 257–271.

Darwin, C.R., 1872. The Expression of the Emotions in Man and Animals. John Murray, London.

Donegan, N.H., Sanislow, C.A., Blumberg, H.P., Fulbright, R.K., Lacadie, C., Skudlarski, P., Gore, J.C., Olson, I.R., McGlashan, T.H., Wexler, B.E., 2003. Amygdala hyperreactivity in borderline personality disorder: implications for emotional dysregulation. Biological Psychiatry 54, 1284–1293.

Ekman, P., Friesen, W.V., 1969. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. Semiotica 1.

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

Ekman, P., Friesen, W.V., 1977. Manual for the Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA.

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

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

Elfenbein, H.A., Ambady, N., 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin 128, 203–235.

Elfenbein, H.A., Ambady, N., 2003. When familiarity breeds accuracy: cultural exposure and facial emotion recognition. Journal of Personality and Social Psychology 85, 276–290.

Erwin, R., Gur, R., Gur, R., Skolnick, B., Mawhinney-Hee, M., Smailis, J., 1992. Facial emotion discrimination: I. Task construction and behavioral findings in normal subjects. Psychiatry Research 42, 231–240.

Golby, A.J., Gabrieli, J.D., Chiao, J.Y., Eberhardt, J.L., 2001. Differential responses in the fusiform region to same-race and other-race faces. Nature Neuroscience 4, 845–850.

Gur, R.C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., Turner, T., Bajcsy, R., Posner, A., Gur, R.E., 2002. A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. Journal of Neuroscience Methods 115, 137–143.

Hare, T.A., Tottenham, N., Davidson, M.C., Glover, G.H., Casey, B.J., 2005. Contributions of amygdala and striatal activity in emotion regulation. Biological Psychiatry 57, 624–632.

Hart, A.J., Whalen, P.J., Shin, L.M., McInerney, S.C., Fischer, H., Rauch, S.L., 2000. Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. Neuroreport 11, 2351–2355.

Herrmann, M.J., Schreppel, T., Jager, D., Koehler, S., Ehlis, A.C., Fallgatter, A.J., 2007. The other-race effect for face perception: an event-related potential study. Journal of Neural Transmission 114 (7), 951–957.

Hugenberg, K., Bodenhausen, G.V., 2003. Facing prejudice: implicit prejudice and the perception of facial threat. Psychological Science 14, 640–643.

Iidaka, T., Ozaki, N., Matsumoto, A., Nogawa, J., Kinoshita, Y., Suzuki, T., Iwata, N., Yamamoto, Y., Okada, T., Sadato, N., 2005. A variant C178T in the regulatory region of the serotonin receptor gene HTR3A modulates neural activation in the human amygdala. Journal of Neuroscience 25, 6460–6466.

Kestenbaum, R., Nelson, C.A., 1990. The recognition and categorization of upright and inverted emotional expressions by 7-month-old infants. Infant Behavior and Development 13, 497–511.

Kraemer, H.C., Periykoil, V.S., Noda, A., 2002. Tutorial in biostatistics: kappa coefficients in medical research. Statistics in Medicine 21, 2109–2129.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lenti, C., Lenti-Boero, D., Giacobbe, A., 1999. Decoding of emotional expressions in children and adolescents. Perceptual and Motor Skills 89, 808–814.

Lieberman, M.D., Hariri, A., Jarcho, J.M., Eisenberger, N.I., Bookheimer, S.Y., 2005. An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. Nature Neuroscience 8, 720–722.

Lobaugh, N.J., Gibson, E., Taylor, M.J., 2006. Children recruit distinct neural systems for implicit emotional face processing. Neuroreport 17, 215–219.

Izard, C.E., 1971. The Face of Emotion. Appleton-century-crofts, New York.

Mandal, M.K., 1987. Decoding of facial emotions, in terms of expressiveness, by schizophrenics and depressives. Psychiatry 50, 371–376.

Mandal, M.K., Harizuka, S., Bhushan, B., Mishra, R.C., 2001. Cultural variation in hemifacial asymmetry of emotion expressions. British Journal of Social Psychology 40, 385–398.

Matsumoto, D., Yoo, S.H., Hirayama, S., Petrova, G., 2005. Development and validation of a measure of display rule knowledge: the display rule assessment inventory. Emotion 5, 23–40.

Phelps, E.A., O'Connor, K.J., Cunningham, W.A., Funayama, E.S., Gatenby, J.C., Gore, J.C., Banaji, M.R., 2000. Performance on indirect measures of race evaluation predicts amygdala activation. Journal of Cognitive Neuroscience 12, 729–738.

Phelps, E.A., Cannistraci, C.J., Cunningham, W.A., 2003. Intact performance on an indirect measure of race bias following amygdala damage. Neuropsychologia 41, 203–208.

Phillips, M.L., Young, A.W., Senior, C., Brammer, M., Andrew, C., Calder, A.J., Bullmore, E.T., Perrett, D.I., Rowland, D., Williams, S.C.R., Gray, J.A., David, A.S., 1997. A Specific substrate for perceiving facial expressions of disgust. Nature 389, 495–498.

Phillips, M.L., Young, A.W., Scott, S.K., Calder, A.J., Andrew, C., Giampietro, V., Williams, S.C., Bullmore, E.T., Brammer, M., Gray, J.A., 1998. Neural responses to facial and vocal expressions of fear and disgust. Proceedings of the Royal Society B: Biological Sciences 265, 1809–1817.

Russell, J.A., 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychological Bulletin 115, 102–141.

Russell, J.A., Bullock, M., 1985. Multidimensional scaling of emotional facial expressions: similarities from preschoolers to adults. Journal of Personality and Social Psychology 48, 1290–1298.

Schlosberg, H., 1954. Three dimensions of emotion. Psychological Review 61, 81–88.

Somerville, L.H., Kim, H., Johnstone, T., Alexander, A.L., Whalen, P.J., 2004. Human amygdala responses during presentation of happy and neutral faces: correlations with state anxiety. Biological Psychiatry 55, 897–903.

Strauss, E., Moscovitch, M., 1981. Perception of facial expressions. Brain and Language 13, 308–332.

Tanaka, J.W., Kiefer, M., Bukach, C.M., 2004. A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study. Cognition 93, B1–9.

Thomas, K.M., Drevets, W.C., Whalen, P.J., Eccard, C.H., Dahl, R.E., Ryan, N.D., Casey, B.J., 2001. Amygdala response to facial expressions in children and adults. Biological Psychiatry 49, 309–316.

Vuilleumier, P., Armory, J.L., Driver, J., Dolan, R.J., 2001. Effects of attention and emotion on face processing in the human brain: an Event-related fMRI study. Neuron 30, 829–841.

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

Winston, J.S., Strange, B.A., O'Doherty, J., Dolan, R.J., 2002. Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nature Neuroscience 5, 277–283.