Valence, arousal, and dominance ratings for facial stimuli

Tina M Sutton, Andrew M Herbert and Dailyn Q Clark

Abstract

A total of 1,363 images from seven sets of facial stimuli were normed using the self-assessment manikin procedure. Each participant provided valence, arousal, and dominance ratings for 120-130 faces displaying various emotional expressions (e.g., happiness, sadness). The current work provides a large database of normed ratings for facial stimuli that complements the existing International Affective Picture System and the Affective Norms for English Words that were developed to provide a normative set of emotional ratings for photographs and words, respectively. This new database will increase experimental control in studies examining the perception, processing, and identification of emotional faces.

During the course of a typical day, we encounter many different stimuli that provoke an emotional response, such as a smile from a friend, a snake in the grass, or a sad story. Emotional stimuli are often characterised by two dimensions: (1) *valence* (ranging from pleasant to unpleasant) and (2) *arousal* (intensity; ranging from calm to excited). Russell and colleagues have found that a two-dimensional model of affective space accounts for 94%-95% of the variance in emotional judgements (Mehrabian & Russell, 1974; Russell, 1980). A third dimension, *dominance* (potency/control; ranging from controlled to in-control), may also be used to discriminate one emotion from another (e.g., "alert" from "surprised" or "angry" from "anxious," Russell & Mehrabian, 1977). However, dominance has not been examined to the same degree as valence and arousal in the emotion literature (Osgood, Suci, & Tannenbaum, 1957). Dominance, or control, accounts for the least variance in affective judgements, and the values provided for this dimension vary more from one participant to the next compared to valence and arousal (Bradley & Lang, 1994).

Researchers have examined the manner in which these dimensions influence how we process emotional words, images, and faces. For example, highly arousing stimuli activate the amygdala, dorsomedial prefrontal cortex (PFC), and ventromedial PFC regardless of stimulus type (Kensinger & Schacter, 2006). Valence has been examined extensively and appears to have stronger effects for emotional pictures compared to emotional words (Bayer & Schacht, 2014). Using a modified photo-word Stroop task, Beall and Herbert (2008) found that emotional faces produced greater Stroop interference effects than emotional words—providing evidence that emotional expressions may be processed more automatically than emotional words (cf. Ovaysikia, Tahir, Chan, & DeSouza, 2010). Dominance has not been examined to the extent that arousal and valence have been studied; however, work by Oosterhof and Todorov (2008) suggests that faces are evaluated on the two dimensions of valence, signalling approach or avoidance, and dominance, signalling the physical strength (or weakness) of an individual's facial cues.

Stimulus sets utilising these three dimensions have been created to provide experimental control for researchers investigating the impact of emotion on cognition. Many of these were created to address perceived faults in other sets of stimuli or to answer a particular research question. Providing consistent measures across different stimulus sets/types can help investigators select emotional stimuli that fit the criteria necessary for their study (e.g., positive and low arousal images). One such database is the

Department of Psychology, Rochester Institute of Technology, Rochester, NY, USA

Corresponding author:
Tina M Sutton, Department of Psychology, Rochester Institute of Technology, George Eastman Hall, 18 Lomb Memorial Drive, Rochester, NY 14623, USA.
Email: tmsgsh@rit.edu

International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2008). IAPS includes images of people (attractive and unattractive; dressed and undressed), animals, houses, objects, etc. that evoke emotions such as joy, sadness, fear, anger, threat, and disgust that have been rated for valence, arousal, and dominance. This stimulus set provides researchers with an abundant set of images that can be matched on valence/arousal/dominance to use in various experiments examining the perception, processing, and identification of emotional information (Bradley & Lang, 2007). Each of the emotional images was normed on the dimensions of valence, arousal, and dominance using the self-assessment manikin (SAM; Lang, 1980). SAM is a graphic figure that ranges from one end of a spectrum (i.e., smiling/happy, excited/wide-eyed, large/dominant) to the other (i.e., frowning/unhappy, relaxed/sleepy, small/non-dominant). Participants select a manikin along the continuum, resulting in a 9-point rating scale for each of the three dimensions. Scores closer to one indicate negative valence, low arousal, and low dominance, and scores closer to nine indicate positive valence, high arousal, and high dominance. IAPS is a widely used database. A recent Google Scholar search for the IAPS technical manual and affective ratings (October 31, 2018) indicated that the IAPS affective ratings of pictures and instruction manual has been cited in 2,801 published papers. Researchers have attempted to replicate and extend the original IAPS ratings. For example, Libkuman, Otani, Kern, Viger, and Novak (2007) collected valence and arousal ratings for 703 of the 716 IAPS images using a Likert-type scale, not the SAM scale procedure. They found similar valence ratings, but the arousal ratings they obtained were lower than those collected using the SAM procedure (Lang et al., 2008). They also extended the original set of ratings by collecting data on the dimensions of surprise, consequentiality, meaningfulness, similarity, distinctiveness, and memorability. Mikels et al. (2005) also extended the IAPS ratings by collecting emotional category data for 203 of the negative images and 187 of the positive images.

Using the same SAM procedure the Affective Norms for English Words (ANEW) was developed by Bradley and Lang to complement the IAPS. ANEW consists of 1,034 words rated for emotional valence, arousal, and dominance (Bradley & Lang, 1999). ANEW provides researchers with a standardised norm for stimulus selection. Many researchers use this database to select words with a particular valence level, arousal level, and/or dominance level. A recent Google Scholar search for Affective norms for English words (ANEW): Instruction manual and affective ratings (October 31, 2018) revealed that the ANEW ratings and instruction manual has been cited in 2,531 publications. Recent work has extended the ANEW database to include 13,915 English lemmas (Warriner, Kuperman, & Brysbaert, 2013). Similar emotion word norms have been developed for languages other than English, such as Dutch (Moors et al., 2013), German (BAWL-R; Võ, Jacobs, & Conrad, 2006), and Polish (NAWL; Riegel et al., 2015).

Research on the perception and recognition of emotional expression has been strongly influenced by the work of Ekman, Friesen and colleagues (e.g., Ekman & Friesen, 1975). They established the notion that there are six universal facial emotions (anger, disgust, fear, happiness, sadness, and surprise) along with the neutral expression. Ekman and colleagues argued that facial expressions of emotion are universal as evidenced by distinct facial musculature used to display each basic emotion and emotion-specific physiological profiles. It is well established that we use facial expressions as a source of nonverbal communication (e.g., Calder & Young, 2005). In addition, emotional expressions assist in social interaction in at least three ways: (1) they provide important information to perceivers that affect behaviour, (2) facial expressions influence responses in social interactions, and (3) produce cooperation among individuals (Keltner, Tracy, Sauter, Cordaro, & McNeil, 2016). Facial stimuli have been used in psychological studies to examine various phenomena such as the influence of emotion on attention (e.g., Mogg & Bradley, 1999; Williams, McGlone, Abbott, & Mattinggley, 2005), cultural differences and/or similarities in emotional perception (e.g., Elfenbein & Ambady, 2002), and the biological substrates of emotional processing (e.g., Kesler-West et al., 2001; Phan, Wager, Taylor, & Liberson, 2002).

Ekman and Friesen compiled one of the first sets of standardised faces, known as the Pictures of Facial Affect (POFA) that is still in use (Ekman, 1976). POFA consists of 110 black and white photographs of 16 Caucasian individuals expressing the six basic emotions along with neutral expressions. Each of the emotion posers was trained using the Facial Action Coding System (FACS) to manipulate certain facial muscles to produce the facial expression of interest. The most recent facial stimulus set published, the Developmental Emotional Faces Stimulus Set or DEFSS, consists of 404 colour photographs of individuals between 8 and 30 years old expressing anger, fear, happiness, and sadness, as well as a neutral expression (Meuwissen, Anderson, & Zelazo, 2017). Facial stimulus sets developed for research differ in many respects. Some used trained actors to display the emotions, such as POFA (Ekman, 1976) and The University of California Davis, Set of Emotional Expressions (UCDSEE; Tracy, Robins, & Schriber, 2009), while others attempted to elicit natural expressions, such as The Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) and The Warsaw Set of Emotional Facial Expression Pictures (WSEFEP; Olszanowski et al., 2015). Databases such as POFA (Ekman, 1976) and the Montreal Set of Facial Displays of Emotion (MSFDE;

Table 1. Listing of the number of faces and emotional expression used in this study along with some pertinent details from the various stimulus sets.

	Angry	Happy	Sad	Neutral	Other	Total	Total Number of models		Races of posers
							Female	Male	
Beall & Herbert	4	5	5	8	–	22	23	13	Caucasian
KDEF	67	67	68	67	201[a]	470	35	35	Caucasian
NimStim	85	128	87	121	–	421	18	25	Various
POFA	17	18	17	14	–	66	8	8	Caucasian
UCDSEE	4	4	4	4	–	16	2	2	Caucasian and West African
WSEFEP	30	30	30	30	–	120	16	14	Caucasian
ONLINE	43	84	45	76	–	248	120	128	Various

KDEF: Karolinska Directed Emotional Faces; POFA: Pictures of Facial Affect; UCDSEE: The University of California Davis, Set of Emotional Expressions; WSEFEP: Warsaw Set of Emotional Facial Expression Pictures.
[a]68: surprise, 67: disgust, and 66: fearful faces from the KDEF stimulus set were included.

Beaupre & Hess, 2005) consist of black and white photographs, while others such as KDEF (Lundqvist et al., 1998) and the NimStim Set of Facial Expressions (NimStim; Tottenham et al., 2009) consist of colour photographs. Facial stimulus sets such as POFA (Ekman, 1976) and WSEFEP (Olszanowski et al., 2015) consist of photographs of individuals of one race or ethnicity, whereas other sets such as NimStim (Tottenham et al., 2009) and UCDSEE (Tracy et al., 2009) provide photographs from individuals of various racial and ethnic backgrounds. Existing stimulus sets also vary on the number of emotions displayed, the number of models photographed, age of the models, eye placement and eye gaze, angle at which the poser was photographed, and validation process (Meuwissen et al., 2017).

The purpose of the rating study described below was to create a tool for researchers similar to the ANEW words and IAPS pictures but for emotional faces. Adolph and Alpers (2010) determined arousal and valence for faces from the KDEF and NimStim face sets. Goeleven, De Raedt, Leyman, and Verschuere (2008) reported a validation study of the KDEF pictures measuring arousal, emotion and intensity for female participants. The present study was intended to sample faces from a broad range of face sets used by researchers in past studies, and provide ratings of valence, arousal, and dominance for the faces that matched certain selection criteria described below.

The starting point was to select all of the photographs displaying anger, happiness, and sadness, as well as neutral expressions from six published sets of facial stimuli and norm them using the SAM procedure utilised in the development of IAPS (Lang et al., 2008) and ANEW (Bradley & Lang, 1999). Pictures of faces expressing other emotions (fear, surprise and disgust) were selected as detailed below to provide researchers with arousal, valence, and dominance ratings for the full range of emotional expressions used in past research. In addition, 248

photographs downloaded from the Internet were also included and normed to provide researchers with another set of stimuli for future studies. A brief description of each of the face sets is provided below.

Previous research has shown the highest inter-rater agreement for happy faces, with decreasing consensus for angry, sad, disgust, surprise and fear. This is evident in Ekman, Sorenson, and Friesen (1969) where the original POFA were shown to different cultural groups, and across all groups the highest agreement was for happy faces (see their Table 1) followed by anger, sadness then the other emotions. The norms provided with the POFA set reflect these differences too, where only happy expressions were recognised as such at over 90% (Ekman, 1976). Table 3 in Ekman (1976) shows the emotion misidentifications of each picture in the POFA, where 11 of 18 happy faces were rated as "happy" by all responders compared to 2 of 18 sad faces, 1 of 18 fear faces, and 5 of 18 angry faces. The Happy advantage is evident in more recent studies and picture sets. For example, Adolph and Alpers (2010) found "decoding accuracy" for the KDEF and NimStim to average 69% to 78% overall. Accuracy for recognising the happy faces was greater than for the other emotions and was higher when the perceived emotion intensity increased. Goeleven and colleagues (2008) validated the Karolinska face set and found highest accuracy for happy faces (over 90%, their Table 1). Angry, disgusted and surprised expressions were recognised at rates between 70% and 79%, with fearful faces recognised below 50% accuracy. The low agreement between raters for emotions other than happy led Beall and Herbert (2008) to only report results for happy, sad, angry and neutral expressions. The decision to focus on the happy, angry, sad and neutral expressions was made because these have been used most often in studies of expression processing and recognition; other expressions have lower inter-rater agreement; and to keep the number of faces to be rated manageable. Nonetheless, a

few other expressions were included to provide a variety of emotions to participants.

Emotional expression stimulus sets

See Table 1 for details of the various face sets and which expressions were used. The Beall and Herbert (2008) Facial Expressions Stimulus Set consists of pictures developed during a pilot study. Each expression poser was given a mirror to help them display each of the emotions. Posers were told to produce each expression with their mouth closed. Multiple colour photographs were taken of each poser, and these were rated as happy, sad, or angry (forced choice) by a total of 34 male and 38 female undergraduates for 1-s stimulus presentations across different rating sessions. The photographs were also rated on a scale from 0 to 6 for each of the three emotions in a self-timed condition. Facial expressions identified consistently as one emotion at a rate of greater than 80% were retained. Beall and Herbert (2008) used a subset of these stimuli in their study where the emotion identification agreement exceeded 95%.

The KDEF (Lundqvist et al., 1998) is a set of 4,900 pictures of facial expressions of emotion. Expression posers were photographed displaying seven different emotions (i.e., afraid, angry, disgusted, happy, neutral, sad, and surprised) from five different angles (i.e., full left profile, half left profile, straight, half right profile, and full right profile). The posers were asked to evoke the emotion being expressed. In addition, they were told to make the expression strong and clear, while maintaining a natural expression of emotion. Only the photos of models looking straight at the camera were selected from this full set of KDEF photos.

The NimStim (Tottenham et al., 2009) stimulus set comprises photos of professional actors. Ten of the models were of African American, 6 were Asian American, 25 were European American and 2 were Latino American. For each emotion, open and closed-mouth expressions were posed, except for surprise, which was only posed with an open mouth. In addition, three versions of happy were posed (closed-mouth, open-mouth, and high arousal open-mouth). The actors were asked to pose the expression "as they saw fit." NimStim photos were selected only for those posers looking straight into the camera to maintain consistency with the photographs in the other stimulus sets.

The POFA (Ekman, 1976) consists of 110 black and white photographs of 16 Caucasian individuals expressing the 6 basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) posed using the FACS (Ekman & Friesen, 1976) to manipulate facial muscles to produce the facial expression of interest.

The UCDSEE (Tracy et al., 2009) is a FACS-verified measure of nine facial expressions posed by four individuals, two Caucasian participants (1 male and 1 female) and two West African (1 male and 1 female) participants. The expressions are posed based on the directed facial action task (Levenson, Carstense, Friesen, & Ekman, 1991).

The WSEFEP (Olszanowski et al., 2015) were created to represent genuine emotion, instead of posed expressions. Participants were asked to focus on a given emotion and then express that emotion to the camera. Each participant was trained to recall an event in which they felt the particular emotion, as well as the physical and psychological sensations produced by that experience.

In addition to these published sets of posed expressions, 248 pictures taken from public domain sources were presented to participants. The faces were selected from a variety of websites including those of news sources. Eligible photos were those where the individual was facing the camera directly (or close to that) and could be cropped out of the entire photograph. Each face photograph had to be 200 pixels wide by 300 pixels tall or larger. The goal was to create a corpus of naturally posed expressions. The majority of the pictures were of Caucasian individuals (166 pictures), 22 of the pictures were individuals of Middle Eastern descent, 21 of the pictures were African American individuals, 19 pictures were individuals of Asian descent, 13 of the pictures were Indian individuals, 5 of the pictures were Hispanic individuals, and 2 of the pictures were of Native Americans.

Method

Participants

Two-hundred and thirty-three participants from Rochester Institute of Technology (RIT) completed the study for course credit or extra credit in a Psychology course. One hundred participants self-identified as female, 132 self-identified as male, and 1 participant self-identified as agender. The average age of the participants was 20 years ($SD = 1.9$). The majority of the sample reported their ethnicity as Caucasian (67%), 11% of the participants were Asian, 9% were African American, 7% were Hispanic, and the remaining 6% were "other." Eight percent of the sample reported that they were deaf or hard-of-hearing.

Materials

A total of 1,363 images were selected from the following sources (permission was obtained when necessary to use the images): (1) Beall and Herbert (2008), (2) KDEF (Lundqvist et al., 1998), (3) NimStim (Tottenham et al., 2009), (4) Online, (5) POFA (Ekman, 1976), (6) UCDSEE (Tracy et al., 2009), and (7) WSEFEP (Olszanowski et al., 2015). The images were then sorted into 11 separate packets of face stimuli such that each packet contained some faces from all of the sources. In addition, the images were randomly divided among the 11 packets and each page showed 10 faces in 2 rows of 5 (landscape orientation).
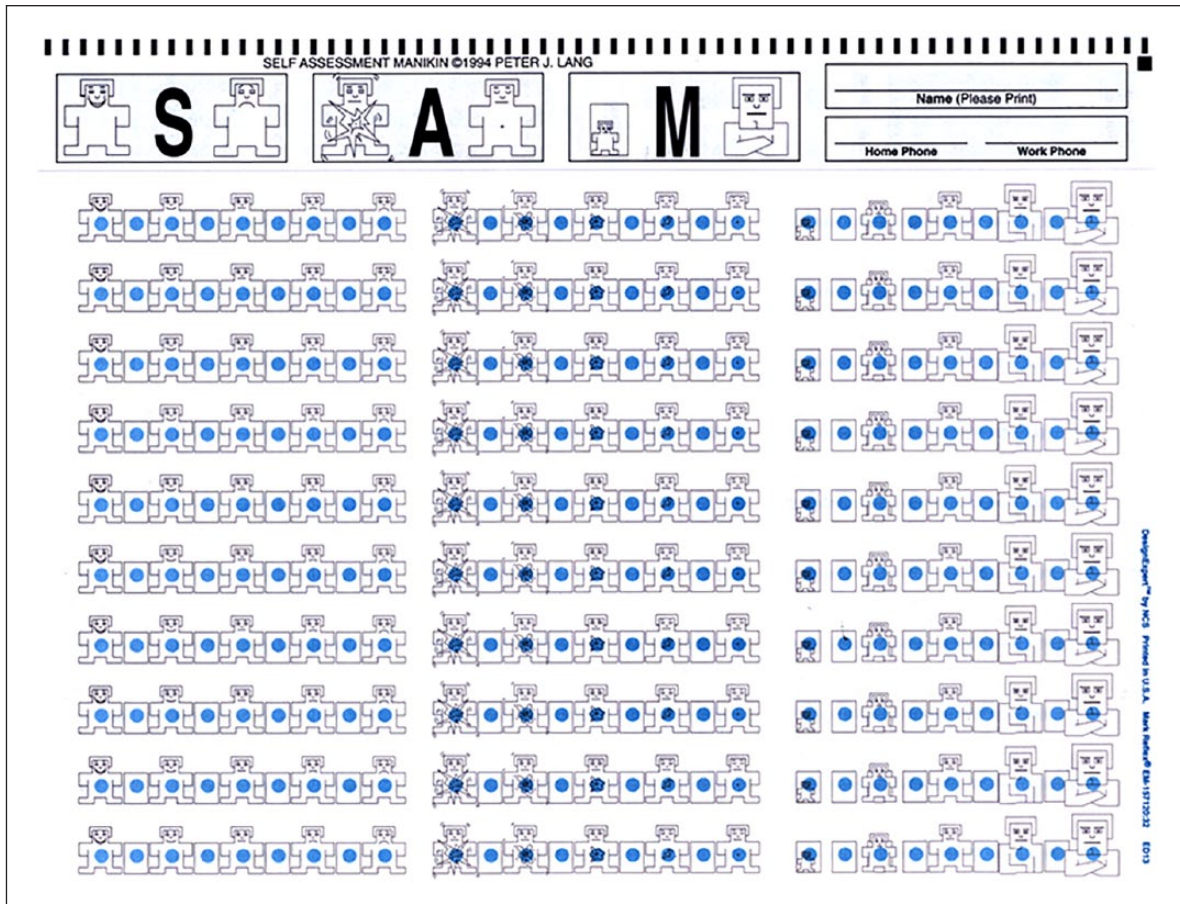
Figure 1. The Self-Assessment Manikin (SAM) developed by Lang (1980) used to rate each of the faces for valence (S), arousal (A), and dominance (M).

A single page contained faces from at least three different stimulus sets with a variety of the emotional expressions, including neutral expressions. All of the images were presented in coloured ink and were 2 inches tall by 1.47 inches wide. Satisfying the various constraints results in four of the packets containing 130 images, six with 120 images, and one packet of 123 images. Raters were randomly assigned a packet to complete.

To rate the faces on the three dimensions of valence, arousal, and dominance, the SAM was used (Lang, 1980). The SAM consists of three graphic figures comprising a 9-point rating scale for each dimension (see Figure 1). The valence dimension ranges from a smiling, happy figure to a frowning, sad figure. The figures for the arousal dimension range from an excited and wide-eyed figure to a relaxed, sleepy figure. The figures for the dominance dimension range from a controlling, large figure to a dominated, small figure. Participants can select among any of the nine boxes along the scale.

Each of the face stimuli were coded by the current researchers using the following scheme: Expression-Gender-Ethnicity-Source-Stimulus Number. The expressions consisted of neutral (NEU), sad (SAD), angry (ANG), fearful (AFR), happy (HAP), surprised (SUR), and disgust (DIS). Gender was coded as male (M) and female (F). Ethnicity was coded as African American (AA), Asian (AS), Caucasian (CA), Hispanic (H), Indian (I), Native American (NA), and Middle Eastern (ME). Source was coded Beall and Herbert (2008) (BEALL&HERBERT), KDEF (Lundqvist et al., 1998), NimStim (Tottenham et al., 2009) (NIMSTIM), Online (ONLINE), POFA (Ekman, 1976) (EKMAN), UCDSEE (Tracy et al., 2009), and the WSEFEP (Olszanowski et al., 2015) (WARSAW). Finally, each stimulus was coded with a number for the current work.

Procedure

The institutional review board (IRB) at the RIT approved the current study. All participants provided written consent prior to starting the experiment. Participants were run in groups ranging from 1 to 28 participants (most groups consisted of 4 or 5 participants). Participants were presented with the SAM figures and each scale was described individually. The researchers described the ends of each continuum and showed participants how to place an X along

any of the nine-points on the scale (instructions were based upon those reported in the ANEW instruction manual). After explaining each scale (valence, arousal, and dominance), participants were shown five sample faces (not included in the experimental packets) to practice using the rating scale. Participants were instructed to rate each picture based on their immediate personal reaction. In addition, they were asked not to compare the pictures to each other, and to rate each picture individually. Finally, they were told that there were no correct or incorrect answers, and that they should rate each picture on all three dimensions. Participants were allowed to ask questions after the practice rating task. At the end of the practice task, participants were given two packets, one with the face stimuli and one with the SAM scales. After completing the ratings, participants were also given a demographics form asking questions about sex, native language, hearing status, age, and handedness. All of the demographics questions were open-ended. The entire session lasted approximately 40 min.

Results

Means and standard deviations for valence, arousal, and dominance were computed for each face in each face set. These are reported in the supplementary material. Each excel workbook includes a worksheet for each face set. In each worksheet, the code we created is matched with the original stimulus set code and the average participant rating of valence, arousal, and dominance for each face is reported. Means and standard deviations were also computed separately for the male and female participants. The ratings for all participants combined, as well as the ratings separated by gender are provided as separate excel workbooks.

We examined the intraclass correlation (ICC) as an estimate of inter-rater reliability (e.g., Shrout & Fleiss, 1979). Participants in the current study received 1 of 11 different packets of stimuli, resulting in a different number of raters for each stimulus; therefore, we computed the ICC for valence, arousal, and dominance for each packet separately. We used ICC (1, k) and reported the mean rating for valence, arousal, and dominance in Table 2. Koo and Li (2016) reported that ICC values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.90 indicate good reliability and values above 0.90 are indicative of excellent reliability. The range of ICC values for valence was 0.979-0.989 and the range of ICC values for arousal was 0.912-0.946, indicating excellent reliability for both dimensions across the raters. The range of ICC values for dominance was 0.793-0.904, indicating good reliability for the dominance dimension.

A Pearson correlation coefficient was computed to compare the average valence and arousal ratings for the NIMSTIM (Tottenham et al., 2009) and KDEF (Lundqvist et al., 1998) faces from the current study that were the same as those rated by participants in Adolph and Alpers (2010). Sixty-seven of the NIMSTIM faces rated by participants for valence and arousal in Adolph and Alpers (2010) were also rated by participants in the current study. A strong positive correlation was obtained for valence, $r(65) = .96$, $p < .01$, and arousal, $r(65) = .71$, $p < .01$. Eighty-eight of the KDEF faces rated by participants for valence and arousal in Adolph and Alpers (2010) were also rated by participants in the current study. A strong positive correlation was obtained for valence, $r(86) = .95$, $p < .01$, and arousal, $r(86) = .69$, $p < .01$.

In addition, a Pearson correlation coefficient was computed to compare the average valence and arousal ratings for 136 KDEF faces from the current study that were rated in Goeleven et al. (2008). Goeleven et al. (2008) only had female participants complete the ratings; therefore, we compared their ratings to those from the female raters in our study. The same SAM scale procedure was used to rate arousal in both studies. A moderate positive correlation was obtained for the arousal measure, $r(134) = .65$, $p < .01$.[1]

A set of analyses was conducted to examine the valence, arousal, and dominance ratings provided for each face expressing anger, sadness, happiness, and neutral from all stimulus sets. A one-way analysis of variance (ANOVA) was conducted to compare the valence values for the four expressions, $F(3, 1,158) = 2,551.06$, $p < .05$, $\eta^2 = 0.87$. Post hoc comparisons using the Tukey's honestly significant difference (HSD) test indicated that the mean valence rating for angry faces ($M = 2.05$, $SD = 0.73$) and sad faces ($M = 2.03$, $SD = 0.75$) were not significantly different from each other; however, the valence ratings for the two negative expressions were significantly different from the valence rating for the positive expression of happiness ($M = 6.52$, $SD = 0.80$). In addition, the valence ratings for all three emotional expressions were significantly different from the valence rating for the neutral expressions ($M = 3.70$, $SD = 0.63$). When examining the SAM valence ratings for angry faces, participants provided ratings on the lower end of the 9-point scale (ranging from 1.0 to 5.0). This was expected given that angry is a negative emotion and lower values indicate negative affect. The same was true for the faces expressing sadness where average SAM ratings ranged from 1.0 to 5.3. Participants provided ratings at mainly at the higher end of the valence scale (range 2.4-8.00) for happy faces.

A one-way ANOVA was conducted to compare the arousal values for the four expressions, $F(3, 1,158) = 323.98$, $p < .05$, $\eta^2 = 0.47$. Post hoc comparisons using the Tukey's HSD test indicated that the arousal ratings for the sad faces ($M = 3.68$, $SD = 1.07$) were significantly different from the arousal ratings for both the angry ($M = 5.02$, $SD = 1.29$) and happy faces ($M = 4.82$, $SD = 1.24$); however, angry and

Table 2. Intraclass correlations for valence, arousal, and dominance separated by packet.

Packet	1	2	3	4	5	6	7	8	9	10	11
Valence	0.988	0.988	0.985	0.989	0.979	0.987	0.989	0.985	0.988	0.982	0.983
Arousal	0.938	0.912	0.942	0.937	0.946	0.933	0.932	0.929	0.918	0.923	0.924
Dominance	0.904	0.812	0.832	0.883	0.840	0.848	0.869	0.793	0.879	0.830	0.847

Intraclass correlations were computed separately for each packet because a different number of raters completed each packet. Twenty-three raters completed packets 8 and 5; 22 raters completed packet 1; 21 raters completed packet 4, 6, 7, 9, and 10; and 20 raters completed packets 2, 3, and 11.

Table 3. Descriptive statistics for the set of ONLINE faces. The average rating across the faces is provided along with the range in parentheses.

Emotional expression	Angry	Sad	Happy	Neutral
Valence	3.1 (1.9–5.0)	2.8 (1.0–5.3)	6.1 (3.6–8.0)	4.1 (2.3–6.7)
Arousal	3.5 (1.7–7.7)	3.6 (1.5–7.1)	4.1 (2.3–7.2)	2.8 (0.8–5.7)
Dominance	5.6 (3.3–7.2)	4.4 (2.1–7.1)	5.5 (3.0–6.9)	5.2 (3.3–7.0)

happy faces did not differ from each other. The arousal ratings for all three emotional expressions were significantly different from the arousal ratings for the neutral expressions ($M = 2.57$, $SD = 0.73$). This indicates that both angry and happy faces are more arousing than sad faces, and that all three of the emotional expressions are more arousing than neutral expressions. The ranges of responses for arousal overlapped across the emotional expressions, ranging from 1.7 to 7.7 for angry, 1.3 to 7.0 for sad, and 2.3 to 8 for happy.

A one-way ANOVA was also conducted to compare the dominance ratings for each of the four facial expressions, $F(3, 1,158) = 332.35$, $p < .05$, $\eta^2 = .01$. Post hoc comparisons using the Tukey's HSD test indicated that the dominance ratings for the sad faces ($M = 3.87$, $SD = 0.83$) were significantly lower than the dominance ratings for the other three expressions, happy ($M = 5.44$, $SD = 0.59$), angry ($M = 5.86$, $SD = 0.89$), and neutral ($M = 5.10$, $SD = 0.76$). In addition, the dominance ratings for the angry faces were significantly higher than the dominance ratings for both the neutral faces and the happy faces. The happy faces were rated higher in dominance than the neutral faces. This suggests that the faces can also be distinguished based on dominance, with sad faces rated the lowest in this scale and angry faces rated the highest. The dominance ratings ranged from 1.5 to 7.8 for angry, 2.1 to 7.1 for sad, and 3 to 6.9 for happy.

Discussion

The purpose of the current work was to provide researchers with a large set of valence, arousal, and dominance ratings for faces to mimic the norms available for words (ANEW; Bradley & Lang, 1999) and images (IAPS; Lang et al., 2008). This set of ratings can be used to provide experimental control while selecting stimuli for experiments examining the perception, processing, and identification of emotional faces. Researchers can select faces matched on valence, arousal, and/or dominance from seven different databases, six previously published sources and a new source, the Online photos. This new set of faces consists of 248 pictures taken from public domain sources (120 females and 128 males) and can be obtained upon request from the authors.

The set of ONLINE faces adds a potential new tool for research on facial expressions. The images were taken from public sources online and comprise unposed photographs of individuals selected based on having happy, sad, angry or neutral expressions. The ratings obtained in the present study are provided in the supplementary materials, and the averages are shown in Table 3. The mean and range for valence, arousal, and dominance for the ONLINE faces match up well with what was found for the other face sets.

We examined the reliability of the ratings provided for valence, arousal, and dominance for the faces rated in the current work by computing the intraclass correlation (ICC) for each dimension across all 11 packets of stimuli. ICC is often used as a measure of interrater reliability. The interrater reliability for valence and arousal were both high, above 0.90, indicative of excellent reliability. The interrater reliability for dominance was lower, but still relatively high (average rating was 0.85), indicative of good interrater reliability.

Adolph and Alpers (2010) compared valence and arousal ratings for faces selected from the KDEF (Lundqvist et al., 1998) and NIMSTIM (Tottenham et al., 2009) datasets. Participants were asked to rate their emotional reaction to the faces using one scale that ranged from "very unpleasant" (1) to "very pleasant" (9) and another scale that ranged from "not at all aroused" (1) to "very aroused" (9). Participants viewed male and female faces in separate blocks. In addition, the faces from the two databases were also presented in separate blocks. In the current work, we required participants to rate their

current level of pleasantness and arousal using the SAM scale procedure. The faces from the various databases were mixed together within the packet and participants viewed both male and female expressions on the same page. A comparison of the valence and arousal ratings for the 88 KDEF faces and 67 NIMSTIM faces rated in both studies revealed a strong positive correlation for valence and a moderately strong correlation for the arousal dimension. This suggests that even though the rating procedures were different across the two studies, participants experienced similar levels of pleasantness and arousal while viewing the faces.

Goeleven et al. (2008) collected emotion, intensity, and arousal ratings for 490 pictures from the KDEF (Lundqvist et al., 1998) dataset. After selecting one of the six basic emotions that matched the expression displayed in the face, participants rated the intensity of that emotion on a scale from 1 ("not at all") to 9 ("completely") and completed the SAM rating scale for arousal (1 = calm and 9 = aroused). Each face was displayed one at a time, and participants rated 39 or 41 pictures. Goeleven et al. (2008) provided the ratings for the 20 best KDEF pictures for each of the basic emotions and neutral expressions in their Appendix. One hundred and thirty six of the faces rated by the participants in Goeleven et al. (2008) were also rated by participants in the current study. Results revealed a moderate positive correlation between the arousal scores obtained by Goeleven et al. (2008) and those collected in the current work.

The stimulus presentation in the current work matched that used for the ANEW words (Bradley & Lang, 1999) and rating norms collected for Dutch words (Moors et al., 2013), with multiple facial expressions presented on each page. Presenting 10 faces per page allowed for more faces to be rated in a single session by each participant. One could argue that the presentation of unfamiliar faces in isolation is largely a laboratory-based phenomenon. Often, we see the faces of strangers in a group, and it is more likely we would interact with familiar individuals when seeing a face by itself. The ratings we obtained were not at the extremes of the scales (as discussed below), suggesting any contrast effects are small if present. Simultaneous presentation could have allowed participants to contrast between sad and happy, or happy and neutral, for example, but these contrasts would also occur in memory with serial presentation. Other studies have had participants provide responses after the faces are presented (Adolph & Alpers, 2010; Goeleven et al., 2008) which requires participants to rate the faces from memory. The fact that similar ratings are obtained across these studies for those faces suggests simultaneous presentation of 10 faces may not affect valence, arousal, and dominance ratings. Participants were instructed to rate each picture individually without comparing them to each other. The compromise of simultaneous versus successive presentation of faces is one of speed versus possible contrast/context effects, and our results suggest any effects of successive or simultaneous presentation are small.

The valence and arousal scores followed a pattern predicted by the Circumplex Model of Affect (Russell, 1980). This model provides a cognitive structure of affect consisting of two dimensions, valence (horizontal axis), and arousal (vertical axis), with affective terms falling in a circular ordering. The negative facial expressions did not differ from each other in terms of valence; however, the mean arousal ratings were different. In addition, the happy facial expressions were significantly different from the two negative facial expressions in terms of valence. Finally, angry faces and happy faces were similar in terms of arousal ratings, but they were significantly different from the sad faces. Spatial models focus on different dimensions of emotional words, mainly pleasure–displeasure (i.e., the tone of the word) and activation (i.e., the sense of mobilisation or energy); however, some models have included dominance-submissiveness as an additional dimension (Russell & Mehrabian, 1977). Russell and Mehrabian (1977) claimed that the dominance-submissiveness factor is a necessary dimension to describe emotion because only dominance makes it possible to distinguish "angry" from "anxious," "alert" from "surprised," "relaxed" from "protected," and "disdainful" from "impotent." The two dimensions of pleasure–displeasure and arousal-sleep could not make these distinctions without the third factor, dominance–submissiveness. It is also important to note that dominance cannot be defined as some combination of pleasure and arousal; therefore, it should be considered an independent dimension. In the current work, sad faces were rated as lower in dominance (feeling less in control when sad) compared to happy, angry, and neutral expressions and angry faces were rated highest in terms of dominance.

It is important to note that the valence, arousal, and dominance ratings were collected from a college-aged sample, and the majority of the sample reported their ethnicity as Caucasian. Research has provided evidence that there is cross-cultural agreement in the recognition of various facial expressions; however, cultural differences in perceived emotional intensity have been reported (e.g., Biehl et al., 1997; Ekman et al., 1987). Specifically, the emotions expressed by individuals in a different culture are rated as less intense than the same emotion displayed by an individual from one's own cultural group. This suggests that the valence ratings may be similar from one culture to another, but the arousal and/or dominance ratings might change if a different sample were tested. Similar to the present work, Roest, Visser, and Zeelenberg (2018) reported high inter-rater reliability for both the valence and arousal dimensions when comparing male and female raters (i.e., valence, $r = .97$ and arousal, $r = .92$), as well as religious and non-religious participants rating general

emotion words (i.e., valence, $r=0.92$ and arousal, $r=0.96$) and taboo words (i.e., taboo general, $r=0.98$ and taboo personal, $r=0.96$) indicating that both the valence and arousal dimensions are judged similarly by different raters. Future work could expand upon the current set of ratings by including additional faces from other published stimulus sets, as well as examining more facial expressions of emotion.

Overall, the supplementary material provides researchers with ratings of valence, arousal, and dominance for a large number of faces from published face sets. Furthermore, we have added a number of photos of unposed emotional faces to the face and emotion researcher's toolkit. The ONLINE faces include exemplars from a variety of ethnicities along with being natural, unposed expressions.

Acknowledgements

Declaration of conflicting interests

Funding

Supplementary material

The Supplementary Material is available at: qjep.sagepub.com

Note

1. In addition to arousal, Goeleven, De Raedt, Leyman, and Verschuere (2008) obtained intensity ratings. After selecting an emotion for the face presented, participants selected an intensity score ranging from (1) *not at all* to (9) *completely*. This intensity rating had no direct equivalent in our study. As can be expected, there was no correlation between the Goeleven et al. (2008) intensity measure and the valence measure used in the current study, $r(134)=0.09$, $p>.05$.

References

Adolph, D., & Alpers, G. W. (2010). Valence and arousal: A comparison of two sets of emotional facial expressions. *American Journal of Psychology*, *123*, 209-219. doi:10.5406/amerjpsyc123.2.0209

Bayer, M., & Schacht, A. (2014). Event-related brain responses to emotional words, pictures, and faces—A cross-domain comparison. *Frontiers in Psychology*, *5*, 1106. doi:10.3389/fpsyg.2014.01106

Beall, P. M., & Herbert, A. M. (2008). The face wins: Stronger automatic processing of affect in facial expressions than words in a modified Stroop task. *Cognition & Emotion*, *22*, 1613-1642. doi:10.1080/02699930801940370

Beaupre, M. G., & Hess, U. (2005). Cross-cultural emotion recognition among Canadian ethnic groups. *Journal of Cross Cultural Psychology*, *36*, 355-570.

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 Nonverbal Behavior*, *21*, 3-21.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*, 49-59.

Calder, A. J., & Young, A. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*, 641-651.

Ekman, P. (1976). *Pictures of Facial Affect. Brochure accompanying the Pictures of Facial Affect developed by Drs. Paul Ekman and Wallace V. Friesen*. San Francisco: Human Interaction Laboratory, UCLA Medical Center.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the Face*. Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*, 56-75.

Ekman, P., Friesen, W. V., O'Sullivan, M., Diacoyanni-Tarlatzis, I., Krause, R., Pitcairn, T., . . . Tzavaras, A. (1987). Universals and cultural differences in judgments of facial expressions of emotion. *Personality and Individual Differences*, *53*, 712-717.

Ekman, P., Sorenson, R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*, 86-88.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*, 203-235. doi:10.1037/0033-2909.128.2.230

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition & Emotion*, *22*, 1094-1118. doi:10.1080/02699930701626582

Keltner, D., Tracy, J., Sauter, D. A., Cordaro, D. C., & McNeil, G. (2016). Expression of emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (pp. 467–482). New York,NY: The Guilford Press.

Kensinger, E. A., & Schacter, D. L. (2006). Processing emotional pictures and words: Effects of valence and arousal. *Cognitive, Affective, and Behavioral Neuroscience*, *6*, 110-126.

Kesler-West, M. L., Andersen, A. H., Smith, C. D., Avison, M. J., Davis, C. E., Kryscio, R. J., & Blonder, L. X. (2001). Neural substrates of facial emotion processing using fMRI. *Cognitive Brain Research*, *11*, 213-226.

Koo, T. K., & Li, M. Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractice Medicine*, *15*, 155-163. doi:10.1016/j.jcm.2016.02.012

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Willimas (Eds.), *Technology in Mental*

Health Care Delivery Systems (pp. 119–137). Norwood, NJ: Ablex Publishing.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A-8*. Gainesville, FL: University of Florida.

Levenson, R. W., Carstense, L. L., Friesen, W. V., & Ekman, P. (1991). Emotion, physiology, and expression in old age. *Psychology of Aging*, *6*, 28-35.

Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., & Novak, N. (2007). Multidimensional normative ratings for the International Affective Picture System. *Behavior Research Methods*, *39*, 326-334. doi:10.3758/BF03193164

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces—KDEF*. CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet, Stockholm, Sweden.

Meuwissen, A. S., Anderson, J. E., & Zelazo, P. D. (2017). The creation and validation of the Developmental Emotional Faces Stimulus Set. *Behavior Research Methods*, *49*, 960-966. doi:10.3758/s13428-016-0756-7

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., & Reuter-Lorenz, P. A. (2005). Emotional category data on images from the International Affective Picture System. *Behavior Research Methods*, *37*, 626-630.

Mogg, K., & Bradley, B. P. (1999). Orienting of attention to threatening facial expressions presented under condition of restricted awareness. *Cognition & Emotion*, *13*, 713-740. doi:10.1080/026999399379050

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A., . . . Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*, 169-177. doi:10.3758/s13428-012-0243-8

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., & Ohme, R. K. (2015). Warsaw set of emotional facial expression pictures: A validation study of facial display photographs. *Frontiers in Psychology*, *5*, 1516. doi:10.3389/fpsyg.2014.01516

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 11087-11092. doi:10.073/pnas.0805664105

Ovaysikia, S., Tahir, K. A., Chan, J. L., & DeSouza, J. F. X. (2010). Word wins over face: Emotional Stroop effect activates the frontal cortical network. *Frontiers in Human Neuroscience*, *4*, 234. doi:10.3389/fnhum.2010.00234

Phan, K. L., Wager, T., Taylor, S. F., & Liberson, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*, *16*, 331-348. doi:10.1006/nimg.2002.1087

Riegel, M., Wierzba, M., Wypych, M., Zurawski, L., Jednoróg, K., Grabowska, A., & Marchewka, A. (2015). Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List—Reloaded (BAWL-R) for Polish. *Behavior Research Methods*, *47*, 1222-1236. doi:10.3758/s13428-014-0552-1

Roest, S. A., Visser, T. A., & Zeelenberg, R. (2018). Dutch taboo norms. *Behavior Research Methods*, *50*, 630-641. doi:10.3758/s13428-017-0890-x

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161-1178.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*, 273-294.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

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

Tracy, J. L., Robins, R. W., & Schriber, R. A. (2009). Development of a FACS-verified set of basic and self-conscious emotion expressions. *Emotion*, *9*, 554-559. doi:10.1037/a0015766

Võ, L.-H., Jacobs, A. M., & Conrad, M. (2006). Cross-validating the Berlin Affective Word List. *Behavior Research Methods*, *38*, 606-609. doi:10.3758/BF03193892

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191-1207. doi:10.3758/s13428-012-1314-x

Williams, M. A., McGlone, F., Abbott, D. F., & Mattinggley, J. B. (2005). Differential amygdala responses to happy and fearful facial expressions depend on selective attention. *NeuroImage*, *24*, 417-425. doi:10.1016/j.neuroimage.2004.08.017