



[Int J Methods Psychiatr Res.](#) 2011 Sep; 20(3): 145–156.

Published online 2011 Aug 24. doi: [10.1002/mpr.343](https://doi.org/10.1002/mpr.343)

PMCID: PMC3342041

NIHMSID: [NIHMS265749](#)

PMID: [22547297](#)

The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS): a new set of children's facial emotion stimuli

[Helen Link Egger](#),¹ [Daniel S. Pine](#),² [Eric Nelson](#),² [Ellen Leibenluft](#),³ [Monique Ernst](#),² [Kenneth E. Towbin](#),⁴ and [Adrian Angold](#)¹

Abstract

With the emergence of new technologies, there has been an explosion of basic and clinical research on the affective and cognitive neuroscience of face processing and emotion perception. Adult emotional face stimuli are commonly used in these studies. For developmental research, there is a need for a validated set of child emotional faces. This paper describes the development of the National Institute of Mental Health Child Emotional Faces Picture Set (NIMH-ChEFS), a relatively large stimulus set with high quality, color images of the emotional faces of children. The set includes 482 photographs of fearful, angry, happy, sad and neutral child faces with two gaze conditions: direct and averted gaze. In this paper we describe the development of the NIMH-ChEFS and data on the set's validity based on ratings by 20 healthy adult raters. Agreement between the *a priori* emotion designation and the raters' labels was high and comparable with values reported for commonly used adult picture sets. Intensity, representativeness, and composite “goodness” ratings are also presented to guide researchers in their choice of specific stimuli for their studies. These data should give researchers confidence in the NIMH-ChEFS's validity for use in affective and social neuroscience research. *Copyright © 2011 John Wiley & Sons, Ltd.*

Keywords: face processing, emotion perception, face stimuli sets, developmental psychopathology, methodology

Introduction

For many years photographs of actors portraying a variety of facial expressions have been used fruitfully as stimuli in psychological experiments. Interest in this area grew from observations on cross-cultural uniformity in the meanings conveyed by these expressions (e.g. Ekman, [1994](#); Ekman and Davidson, [1994](#)). With the more recent emergence of new neuroscience technologies, such as functional magnetic resonance imaging (fMRI), there has been an explosion of basic and clinical research on the affective and cognitive neuroscience of face processing and emotion per-



ception (e.g. Adolphs *et al.*, [1994](#), [2002](#); Nelson *et al.*, [2003](#)). This neuroscience work suggests that, in humans, facial emotion displays engage a dedicated neural architecture that also is engaged by emotionally-salient cues in many other mammal species (Ekman and Davidson, [1994](#)).

Work on face-emotion displays in psychology and neuroscience benefits from the availability of stimulus sets that generate reliable and valid ratings for a range of face emotions. These stimulus sets ideally should include photographs from a large number of individuals because fMRI work suggests that key brain structures habituate to repeated presentations of the same stimulus (Breiter *et al.*, [1996](#); Thomas *et al.*, [2001](#)). The stimuli sets must also display high quality images.

A number of face stimuli sets of adult faces, including the venerable black and white Ekman set produced in the 1970s, the Japanese and Caucasian Facial Expressions of Emotion (JACFEE), and the more recent color NimStim set of Facial Expression, are publically available for use in psychology and neuroscience research (Ekman, [1976](#); Mazurski and Bond, [1993](#); Evalynn *et al.*, [1996](#); Biehl *et al.*, [1997](#); Wang and Markham, [1999](#); Gur *et al.*, [2002](#); Palermo and Coltheart, [2004](#); Calvo and Lundqvist, [2008](#); Tottenham *et al.*, [2009](#)). These stimuli sets vary in the number of emotional expressions included (out of six basic emotions: anger, afraid, sad, happy, disgust, surprise), the inclusion of neutral or calm stimuli, the ethnicity/racial background of actors, the digital quality and digital standardization of the photographs, the number of pictures per condition, posing characteristics (e.g. open or closed mouth [Tottenham *et al.*, [2009](#)]; varied profiles (Gur *et al.*, [2002](#); Calvo and Nummenmaa, [2009](#)), and the availability of reliability and validity data on the individual pictures, condition subsets, or overall set. The primary measure of the validity of face picture sets has been the degree of agreement between the images *a priori* emotion designation and the raters' identification of emotion type. Associated features including intensity of the emotion, reaction time to make ratings, and gender and cultural differences have also been examined (Ekman, [1976](#); Mazurski and Bond, [1993](#); Biehl *et al.*, [1997](#); Wang and Markham, [1999](#); Gur *et al.*, [2002](#); Palermo and Coltheart, [2004](#); Calvo and Lundqvist, [2008](#); Tottenham *et al.*, [2009](#)). For most picture sets, overall mean proportion correct reported ranges from 70–88%.

Face stimuli have been extensively used in developmental psychology and psychopathology, most notably in studies of face and emotion processing dysfunctions and their associated neural substrates, in children with various mental illnesses (Adolphs *et al.*, [2001](#); Pelphery *et al.*, [2004](#); Dalton *et al.*, [2005](#); Guyer *et al.*, [2007](#); Brotman *et al.*, [2008](#), [2010](#); Perlman *et al.*, [2009](#)). Most studies of children's responses to facial expression stimuli have relied upon children's responses to adult faces. Stimuli sets of child faces are needed to explore children's responses to peer emotional faces and to test whether there are differences in children's and adults' responses to adult and child faces. Research in the neural substrates of pediatric emotion perception and processing would benefit from a large, high-quality, digital, standardized set of facial emotion stimuli of children. To our knowledge only one available picture set includes children (only three children and a total of 42 pictures) (Mazurski and Bond, [1993](#)). To address the need for a set of facial expressions of children, the Emotion and Development Branch in the National Institute of Mental Health (NIMH) Division of Intramural Research Programs created the NIMH Child Emotional Faces Picture Set (NIMH-ChEFS), a new set of high resolution facial emotion stimuli of children ages 10 to 17 years old. The pictures are color, high quality digital images that capitalize on advances in digital photography and image processing. The NIMH-ChEFS is freely available to the scientific











community for use. In this paper we describe the creation of the stimulus set and present the results of an initial evaluation of this new picture set based on the scoring of the pictures by 20 healthy adult raters.

Methods

Development of the stimuli set

The NIMH-ChEFS was created through a collaborative endeavor between a neuroscience research group at the NIMH and the Imagination Stage, a local children's theater group based in Bethesda, Maryland, a town within the greater Washington DC metropolitan area. Through a series of meetings, the research and theater group leaders shaped the goals of the project, which was to obtain a set of high-quality photographs of children posing various facial emotion displays. Photographs were taken by a professional photographer in a series of two-hour sessions conducted over a two week period in 2004. A theater teacher and a neuroscientist from the NIH research group staffed these photography sessions. Child actors were enrolled in classes at the children's theater. The teachers at the theater company selected children who were thought to be most effective at portraying emotions. The agreement to obtain the stimulus set was a business arrangement with the local theatre group and parents. Parent permission and actor assent to make the pictures publicly available for researchers and to be reproduced in scientific publications and presentations were obtained by contractual arrangement. Children were given professional-grade digital "head-shot" photographs in lieu of payment for participation. The arrangements were reviewed with NIMH IRB which considered this to be a non-research activity outside IRB purview.

Procedures for posing specific face-emotion displays were extracted from details provided by Ekman and Friesen ([1975](#)). The team of neuroscientists reviewed these details with the teachers at the children's theater group, who, in turn, reviewed the procedures with each child actor. Photographs were acquired in a separate session. The theater coach instructed each child actor to pose a specific face emotion display. While the actor posed this emotion, five-to-10 photographs were acquired for each of the gaze conditions (direct and averted). Pictures were then taken for each subsequent emotion, until five-to-10 photographs had been acquired for five emotions (happy, sad, fearful, angry, and neutral) and two conditions (direct and averted). Photographs were acquired at three angles: one directly in front of the actor and one on each side of the actor. Representative images from the picture set are shown in [Figure 1](#).

Condition	Relative rating	Image	Agreement (N=20)	Intensity (SD)	Representativeness (SD)
Afraid averted	Highest rated		20	92 (10)	90 (13)
	Lowest rated		12	1 (59)	2 (36)
Angry direct	Highest rated		20	82 (14)	86 (15)
	Lowest rated		7	-9 (54)	-1 (54)
Happy direct	Highest rated		20	90 (13)	87 (14)
	Lowest rated		17	40 (41)	51 (50)
Neutral averted	Highest rated		20	73 (17)	NA
	Lowest rated		7	-17 (58)	NA
Sad direct	Highest rated		20	90 (10)	80 (22)
	Lowest rated		0	-45 (23)	-42 (21)

[Figure 1](#)

Representative images from the picture set; SD, standard deviation.

Description of the stimulus set

The research team at NIMH then reviewed the pictures and selected the best picture for each child for each emotion and each condition (Table 1). Only front facing pictures are used in the picture set.

Table 1

Ages of individual child actors by stimulus number

F10	14	M10	12
F11	14	M11	11
F12	12	M12	11
F13	14	M13	14
F14	15	M14	15
F15	15	M15	13
F16	12	M16	13
F18	16	M17	10
F19	13	M18	15
F20	14	M19	12
F21	17	M20	13
F22	14		
F23	11		
F24	14		
F25	13		
F26	15		
F27	15		
F28	16		
F29	Unknown		
F30	13		
F31	15		
F32	11		
F33	14		
F34	14		
F35	14		
F36	14		
F37	12		
F38	10		
F39	10		
F40	12		

Note: There is no stimulus F17. Age of F29 is not known.

The original picture set includes 534 pictures (341 girls, 193 boys). Total number of pictures per emotion and gaze condition can be found in Table [2](#). There are 39 girls and 20 boys in the picture set (total $N = 59$). The child actors ranged in age from 10 to 17 years old with a mean age of 13.6 years old (no difference in mean age for boys and girls). Table [1](#) shows the age of the each child actor by stimulus number. Information on the child actors' ethnicity was not obtained. Based on appearance, most of the actors are Caucasians with four girls and one boy appearing non-Caucasian. For the majority of children, the set includes 10 pictures for each child (five emotions \times two conditions). Six female subjects (F24–F29) and one male subject (M13) have incomplete sets of photographs. As described in the results section, we set a validity cutpoint for inclusion in the recommended set of 15/20 or more raters correctly identifying the intended emotion. This cutpoint excludes 52 pictures for a final set of 482 pictures.

Table 2

Number, agreement, intensity, representativeness, and goodness ratings, by condition

Stimulus type	Total picture <i>N</i>	Observations <i>N</i>	Percent agreement		Intensity	Representativeness	
			for original <i>N</i>	<i>N</i> with < 75% agreement			<i>N</i> with 75% + agreement
Afraid averted	56	1100	93	3	53	55	51
Afraid direct	52	1040	95	3	49	60	54
Angry averted	52	1040	90	7	45	59	59
Angry direct	52	1040	93	3	49	59	59
Happy averted	56	1120	99	0	56	67	76
Happy direct	52	1040	99	0	52	72	82
Neutral averted	52	1040	83	9	43	—	48
Neutral direct	59	1180	93	4	55	—	65
Sad averted	53	1060	81	11	42	48	49
Sad direct	50	1020	78	12	38	51	55
Total	534	10,680	90.4	52	482	60	61

Note: Intensity and representativeness scales range from -100 to 100.

The NIMH-ChEFS can be downloaded for use in research at the following website (<http://devepi.duhs.duke.edu/instruments.html>) or by contacting Dr Pine's research group at NIMH. There is no cost for obtaining or using these images.

Evaluation of the stimulus set

The research group at the Center for Developmental Epidemiology (CDE) at Duke University Medical Center undertook an evaluation of the NIMH-ChEFS in order to select the best stimuli for an fMRI paradigm in an on-going longitudinal imaging study of early childhood anxiety disorders (RO1 [MH081025](#)). After a review of previous evaluations of face stimuli sets, CDE investigators (AA and HE) developed the following rating paradigm and study.

The raters The stimuli were evaluated by a convenience sample of 20 volunteers, all faculty and staff working in the CDE at Duke University Medical Center. The Duke IRB approved development of methodologies in the study for which these images were to be used. Those who completed the evaluation were rewarded with a pizza lunch paid for by CDE faculty (AA and E. Jane Costello). Raters ranged in age from 22 to 70 (mean age 38.3); 13 were women and seven were men. Five have doctoral degrees; one has a master's degree, and the rest have undergraduate college degrees.

The evaluation paradigm A CDE software developer (Tim Blitchington) created a web-based flash program to present the stimuli for rating. The program was hosted on the CDE intranet and was only accessible to staff at the CDE. Evaluations were completed on each rater's usual computer over a secure link to the Center's network. Upon signing in to complete the evaluation, a startup screen appeared in which the rater entered information about his/her sex, age and educational level. Next a description of the rating tasks was presented, after which the evaluator was shown the first stimulus.

Figure [2](#) is a screen shot of an evaluation page completed for each image. The stimulus picture fills the left half of the screen frame. On the right are three tasks for the rater to complete. First, the rater was told to select from a list the emotion that he or she feels the picture represents. The list includes the five stimulus emotion types (afraid, angry, happy, neutral, sad). Second, the rater was instructed to set a slider labeled "intensity of the emotion." The slider ranged from mild to strong. This judgment was to be made on the basis of the degree of emotion being expressed, regardless of whether the rater was confident that the he or she had correctly classified the emotion type. So, for instance, in the case of a stimulus meant to represent sadness, a rater might be clear that some strong emotion was being expressed (and therefore rate it high on intensity), but may not have had confidence that the stimulus expressed sadness or anger and might not have selected sadness as the emotion type.



[Figure 2](#)

Screen shot of an evaluation page.

Third, the rater was told to rate how accurately the picture represented the emotion selected using a second slider, labeled “representativeness,” anchored by the words “poorly” and “very well.” In this case, the raters had been instructed to consider the degree to which the stimulus appeared to them to be a *good* representation of the emotion that they thought it represented. So, in the example described earlier, the high intensity rating would have been accompanied by a low representativeness rating because the rater could not tell whether sadness or anger was being expressed.

The computer program administering the stimuli randomized the order of presentation for each rater. Ratings were saved each time the rater clicked through to the next stimulus.

Raters were allowed to proceed at their own paces and move on to the next stimulus whenever they were ready. They were also allowed to take any breaks they wanted and to complete their rating over multiple sittings. Individual raters took between one and a half to four hours to complete all 534 ratings.

From these ratings, we created a SAS data set with 32,580 rating observations (543 pictures \times 20 raters \times three ratings per picture). The probability of obtaining 15/20 agreement if there really was no signal (i.e. random distribution expected) would be 1.4×10^{-8} (i.e. 1.4 in 100 million).

Analytic approach We scored each stimulus image on a number of dimensions:

1. Agreement/disagreement rate – the number of instances in which a rater considered an image to be an expression of an emotion different from that which it was intended to convey.

2. Intensity – In cases where the rater and the *a priori* designation of the stimulus agreed on the emotion portrayed, the intensity score was simply that recorded for that stimulus. However, when the *a priori* and rated emotions differed we considered that the intensity score should be penalized because a rating of high intensity of a different emotion was evidence that the stimulus had not functioned as intended. In these cases we multiplied the intensity score by -1 (so that an intensity score of 37 would become -37). The final scale, therefore, had a possible range of -100 to 100 . The neutral stimuli were an exception here. Neutral stimuli are intended to avoid high (or any) emotional intensity. We expected that “good” neutral stimulus would be low intensity. However, we found that the mean intensity rating for neutral stimuli was 43, higher than that for the sad stimuli (35). This led us to question study participants about how they had approached the intensity ratings for neutral stimuli. We found that most rated a stimulus that they thought was “very neutral” as having relatively *high* “neutral intensity.” However, some other raters adopted the reviewer's position that neutrality implied low intensity. We, therefore, concluded that where neutral stimuli were concerned, we had not adequately specified the meaning of the intensity ratings, with the result that they were not really interpretable. In this paper, we provide raw data on individual neutral stimuli but not composite intensity data for the neutral stimuli because we concluded that the results were not meaningful.
3. Representativeness – The raw representativeness scores were treated in the same way as the intensity scores.

We then examined the mean levels of these four parameters by emotion condition and gaze type. We also examined the mean proportion correct for each emotion condition, as well as the variety and frequency of mismatches between the different emotion types. From these data, we calculated an overall kappa to estimate the concordance between the intended emotion and the raters' labels. Based on Wagner's suggestion that confounding of recognition accuracy and frequency of usage of each response category may inflate accuracy rates, we also calculated differential accuracy (unbiased hit rate) for each emotion condition which is an estimate of the joint probability both that a stimulus is correctly identified (given that it is present) and that a response is correctly used (given that it is used) (Wagner, [1993](#), p. 16).

For the individual images we also generated a composite goodness score to facilitate researchers in choosing individual images for their paradigms. Goodness was computed as the product of the intensity and accuracy scores divided by 100 to produce a compound measure of the overall “strength” of the stimulus (with a possible range from -100 to 100). Where goodness and accuracy were both negative, the goodness score was multiplied by -1 to maintain its utility as an overall measure of the degree to which a stimulus performed its intended function. Since there were no meaningful intensity ratings for the neutral stimuli, there could also be no meaningful goodness scores. For most purposes, the “best” stimuli will be those with low disagreement rates and high levels of intensity, representativeness and, therefore, goodness.

SAS 9.2 was used for all analyses.

Results

Table 2 includes percent agreement, intensity, and representativeness ratings, overall and for each emotion by gaze condition.

Agreement with *a priori* classification, overall

The mean disagreement rate was 9.4% with a range from 3.1% to 18.9% and an interquartile range from 7.9% to 10.7%. The 90th percentile lay at 12.6%. For almost half of the stimuli (49.5%) there was complete agreement among all the raters and the *a priori* emotion classification. In a further 17.5% of cases one rater dissented, while 9.8% of stimuli were agreed upon by 18 of the 20 raters. Thus for 76.7% of the stimuli 18 or more raters agreed with the stimulus' original emotion designation. For 90.4% of the stimuli, at least 15 (75%) of the raters agreed with the *a priori* emotion designation. Based on standards used in prior work, we set a cutpoint requiring at least three quarters of the raters to have agreed with the *a priori* classification for the inclusion of stimuli in the picture set. Thus, 9.6% of the stimuli (52 images) are considered to have poor support and thus should probably not be used. We have defined the 482 picture set consisting of images meeting this cutpoint as the acceptable picture set (see supplemental Table S1 for included and excluded pictures).

Agreement by emotion condition

There were quite substantial differences among the different emotion types, as shown in Table 2. Whereas all the happy faces were identified as such by at least 15 raters (and, actually, by at least 18), only 76% of the sad direct gaze faces were identified by that number. Thus, none of the happy faces were excluded from the picture set while six of the afraid faces (three averted, three direct), 10 of the angry faces (seven averted, three direct), 13 of the neutral faces (nine averted, four direct), and 23 of the sad faces (11 averted, 12 straight) had less than 75% agreement. Overall 30 averted gaze pictures were excluded versus 22 direct gaze pictures. However, for the sad group which had the most unacceptable pictures, there remained 38 recommended stimuli.

Table 3 summarizes the agreement proportion across the five emotion conditions, as well as measures accounting for chance agreement between ratings and intended emotion (kappa for whole sample and index of accuracy for each emotion type). The top half of Table 3 includes data for the entire picture set and the bottom half of Table 3 are data for the validated set. The agreement proportions for the entire set range from 0.80 to 0.99 with accuracy index scores ranging from 0.67 to 0.99. In the validated set all of the proportions of agreement are greater than 0.90, where as the index of accuracy range from 0.84 to 0.99.

Table 3

Percent agreement and disagreement for each emotion type and index of accuracy for each emotion type, for entire set and for validated set

Photograph	Emotion identified by raters				
	Afraid	Angry	Happy	Neutral	Sad
<i>Entire set of 534 pictures (10,680 observations), Kappa = 0.86</i>					
Afraid	2018 ^a	56	2	12	52
	0.94 ^a	0.26	0.0009	0.06	0.02
	0.86 ^a				
Angry	44	1893 ^a	5	51	87
	0.02	0.91 ^a	0.024	0.025	0.04
		0.77 ^a			
Happy	4	2	2147 ^a	0	7
	0.002	0.002	0.99 ^a		0.003
			0.99 ^a		
Neutral	61	75	7	1961 ^a	116
	0.03	0.03	0.003	0.88 ^a	0.05
				0.81 ^a	
Sad	97	200	6	123	1654 ^a
	0.05	0.09	0.003	0.06	0.80 ^a
					0.67 ^a
<i>Acceptable ^b set of 482 pictures (9640 observations), Kappa = 0.94</i>					
Afraid	1957 ^a	40	2	10	31
	0.96 ^a	0.02	0.001	0.005	0.02
	0.90 ^a				
Angry	33	1785 ^a	0	24	38
	0.02	0.95 ^a		0.01	0.02
		0.89 ^a			
Happy	4	2	2147 ^a	0	7
	0.002	0.001	0.99 ^a		0.003
			0.99 ^a		
Neutral	48	44	7	1804 ^a	57
	0.02	0.02	0.004	0.92 ^a	0.02

Note: Cells show: *N* observations, agreement proportion, index of accuracy for target emotion.

^aTarget emotion.

^bImages meet validity criterion (15 or more/20 raters agreed on label).

Table 3 also shows differences in emotion misattribution by emotion type. Of the 5.7% mislabeled afraid pictures, 45% were called angry and 42% were called sad, with only 10% being labeled neutral and less than 2% labeled happy. Half of the mislabeled angry pictures were called sad, with 23% choosing afraid and 27% choosing neutral. For the 0.6% mislabeled happy pictures, none labeled them as neutral. Of the mislabeled neutrals 45% were identified as sad, with another 24% labeled as afraid and 29% labeled as angry. Lastly, for the sad faces, 47% were named angry, 23% called afraid, and 28% called neutral. Few of the non-happy faces were identified as happy.

For all emotion and gaze conditions, 90.3% of the girl pictures and 90.2% of the boy pictures were identified correctly by 15 or more of the raters. Because the complete picture set has more girl pictures than boy pictures, 63.9% the acceptable set of 482 pictures are of girls.

Intensity and representativeness ratings

The mean intensity and representativeness ratings by stimulus type are presented in Table 2. Intensity and representativeness were very highly correlated (Pearson $r = 0.85$). The mean intensity score across the fearful, angry, happy and sad stimuli was 60 (standard deviation [SD] = 33); the mean representativeness across all five emotion conditions was 61 (SD = 31).

It was immediately apparent that some stimulus types attracted higher ratings than others, with an overall mean intensity score of 72 for happy direct gaze faces compared with 48 for sad averted faces. The pictures of happy faces were found to be more representative of happiness than neutral averted gaze pictures were of neutrality. However, in every case, there was a reasonably even distribution of stimuli across the available ranges. There were no significant differences in ratings between the direct and averted gaze conditions; for example, mean intensity for the averted gaze stimuli was 57.3 whereas the mean intensity for the direct gaze stimuli was 60.5.

There were significant differences in ratings by sex of face picture for the negative emotions. For the afraid pictures, the girl faces were rated as significantly more intense (59 versus 50; $p < 0.0001$) and more representative of fear (54 versus 43; $p < 0.0001$) than the boy faces. The girl sad faces were rated significantly more representative than the boy sad faces (40 versus 32; $p = 0.002$), and there was a trend suggesting that the girl sad faces were also more intense (37 versus 33; $p = 0.09$). However, the boy angry faces were rated as more intense than the angry girl faces (60 versus 51; $p < 0.0001$). There were no significant differences in the intensity or representativeness ratings for the happy or neutral faces.

Validity of individual images

A table of agreement, intensity, representativeness and “goodness” rating for each of the 534 images, as well as a list of the pictures excluded, can be found in supplemental Table S1. Given the enormous size of Table S1, it is to be found in its entirety online at [insert where supplemental in-

formation is hosted]. Here we present, in Figure 1, examples of the highest and lowest scores for each emotion condition (three direct gaze, two averted gaze), along with the specific images. For example, the highest rated afraid, averted gaze picture was identified as showing fear by all 20 raters. It had a very high intensity score with a very low SD; it was also regarded as being an excellent portrayal of fear. Conversely, the poorest sad, averted gaze picture was rated as showing sadness by only one rater. Two others rated it as showing fear, and the remaining 17 rated it as showing anger. Its overall intensity, representativeness, and goodness scores were therefore negative. As noted earlier, there were significant differences in the range of goodness between the best and worst pictures: the worst rated happy, straight gaze pictures had a goodness score of 31 while the worst averted gaze picture had a goodness score of -42. Based on the aims of a particular study and the experimental paradigm, researchers will be able to use these four variables, alone or in combination, to select the best images from this picture set for their particular needs.

Discussion

We set out to create and present an initial evaluation of the NIMH-ChEFS, a new set of child-based, high resolution images of fearful, angry, happy, sad and neutral faces with two conditions: direct and averted gaze. To test the validity of the stimuli, we used ratings by a convenience sample of 20 raters who were asked to identify the emotion being expressed and then rate its intensity and representativeness. Approximately 10% of the stimuli from the original picture set were invalid, in that fewer than three quarters of the raters agreed on its *a priori* categorization (i.e. fewer than 15/20 raters correctly labeled the intended emotion). For both the entire picture set ($N = 534$) and for the acceptable set ($N = 482$), measures of validity were high. Across all of these measures, agreements on the level of the individual picture, emotion condition, emotion by gaze type, and overall picture set were high, comparing favorably with values reported for commonly used adult picture sets.

Our primary measure of validity was agreement between the *a priori* emotion and the raters' label, calculated as percent agreement or proportional agreement. We also calculated Wagner's index of accuracy for each emotion condition and Kappas for the entire picture set in order to account for the 1/5 chance that label and picture will be matched randomly. The overall percent agreement for the entire picture set was 90.4% with a Kappa of 0.86; for the acceptable pictures set the overall percent agreement was 94.8% with a Kappa of 0.94. The mean proportion correct for adult picture sets are comparable (e.g. the NimStim picture set [0.81], the Ekman Pictures of Facial Affect [0.88], and the JACFEE set [0.74], the Karolinska Directed Emotion Faces [0.89] [Ekman, 1976; Biehl *et al.*, 1997; Calvo and Lundqvist, 2008; Tottenham *et al.*, 2009]). Wang and Markham (1999) set a 70% agreement cutpoint for inclusion of pictures in their set of Chinese faces. In a validity study containing images from five picture sets including the NimStim and the Ekman Pictures of Facial Affect, Palermo and Coltheart reported an overall labeling accuracy of 76.4%. Lastly, the overall Kappa for the NimStim set was 0.79.

As in studies of other picture sets, happiness had the highest agreement ratings (e.g. Biehl *et al.*, 1997; Calvo and Lundqvist, 2008; Ekman, 1976; Gao and Maurer, 2009; Tottenham *et al.*, 2009; Wang and Markham, 1999) with the negative affects of fear, anger and sadness having relatively lower agreement rates. In this picture set, sadness was the least accurately identified emotion,

whereas fear has been the least accurately identified emotion in previous studies of adult face stimulus sets (e.g. Ekman, [1976](#); Calvo and Lundqvist, [2008](#); Tottenham *et al.*, [2009](#)). We excluded about a quarter of the sad stimuli from the acceptable data set. Whether this difference is due to the child actors' skill at expressing sadness or adults' difficulty recognizing children's sadness is not clear (the second possibility is testable). Nonetheless, as noted earlier, our accuracy index of 0.67 for the whole set and 0.84 for the accepted picture set are comparable to those reported for adult picture sets (e.g. NimStim kappas for sad images [closed mouth: 0.83, open mouth: 0.60]). All of the validity measures for the 10 emotion-by-gaze conditions in the entire picture set were high and even higher for the acceptable picture set.

Even with elimination of about 10% of the pictures for decreased accuracy, the acceptable set has between 38 and 53 stimuli in each of the emotion by gaze direction categories. Examination of the ratings of stimulus intensity and representativeness showed that in every category there was a wide range of stimulus “strengths,” ranging from the subtle to the blindingly obvious. We regard this as being a useful property, because different experimental paradigms call for stimuli of different “strength.” We also calculated a composite “goodness” rating for each of the images. All of these ratings, for the individual pictures and for the stimuli of a given emotion type, can be used by researchers in selection of specific stimuli for their studies. No sex difference was found in the accuracy of the recognition of any emotions as has been noted in at least one previous study (Palermo and Coltheart, [2004](#)). However, we did find that the afraid and sad girl faces were rated as more intense than the boy faces, while the angry boy faces were rated as more intense than the girl faces, differences which merit further exploration.

Limitations

The forced choice design of the evaluation study

Some investigators might object to the omission of choices for surprise or disgust in our rating paradigm. Since fearful facial expressions are more similar to surprise than to any of the other included expressions, if a stimulus actually looked like surprise, it would naturally have been classified as fear because it did not resemble sadness, happiness, anger or neutrality. In other words, intelligent guesswork could have increased the number of correct attributions.

We considered including surprise and disgust in the available choices in order to obviate this possibility. However, the inclusion of these choices would likely have led a rater to expect that surprise or disgust were included among the face emotions to be displayed when none were. Such an expectation would result in the raters thinking that they must be failing to identify surprise and disgust stimuli and lead them to lowering their thresholds for identifying them, leading, in turn, to artifactual disagreement with the *a priori* designation. The developers of the NimStim picture set included a “none of the above” option in their evaluation measure (Tottenham *et al.*, [2009](#)). We did not include this option because we felt that our representativeness scale would capture the rater's impression that the depicted emotion was a poor representation.

Adult raters

One may also ask why we conducted our evaluation with adults rather than children. The purpose of this study was to test the validity of these stimuli as representations of specific emotional expressions. We purposefully chose a sample of adults, all of whom work with children in our laboratory, who would be well equipped to identify emotional expressions. Certainly, it is worth testing whether children's ratings of the stimuli would have the same level of agreement and similar ratings of intensity and representativeness as adults' ratings. Yet, the purpose of this study is to assess the validity of this picture set as representations of specific emotional expressions. Our data suggest that this new picture set is as valid as other picture sets currently being used in psychology, developmental psychopathology, and neuroscience research. Having demonstrated that these pictures *are* valid representations of specific emotional facial expression, we can now consider children's ratings and explore differences by age and diagnosis. Our rating program can easily be adapted for use in a study of children's ratings of the stimulus set. The paradigm would have to be shortened significantly because we would not expect that children, even adolescents, would have the fortitude to rate over 500 pictures. Paradigms would also need to be modified based on developmental and cognitive differences across childhood. The CDE will share our rating program with other research groups interested in conducting similar evaluations at their sites (please contact the first author by email to obtain the program). It would also be valuable to conduct similar validity studies within different ethnic/racial groups, as has been done with adult picture sets (e.g. Biehl *et al.*, [1997](#)) to assess cultural bias in these pictures.

We certainly are aware of the limitations of using this new picture set without developmentally specific validity and intensity rating data and caution researchers to consider this limitation when using these norms to guide research with developmental samples. Of course, use of this picture set in developmental neuroscience studies outside of the NIMH and CDE laboratories will contribute to the on-going evaluation of their validity and usefulness. Our group is currently collecting eye tracking data using the NIMH-ChEFS and the adult NimStim pictures with a community sample of 500 preschool children (ages 2–5), half of whom meet criteria for an anxiety disorder and half who do not. We are also conducting an fMRI study with a subset of these children at ages 6–8 years old using both picture sets. While not a substitute for a validity experiment with children, these data will provide an opportunity to examine the similarities and differences in children's responses to specific emotional faces of adults and children and consider differences by age and by diagnosis. In [Table 1](#) we report the age of each child actor so that researchers can restrict the age range of stimuli to match the age range of an experimental sample or control for age across emotion conditions.

While this stimulus set was developed for pediatric affective neuroscience research, there is no reason to think that their use would be confined to children. Exploration of differences in how adults respond to children's emotional faces, compared to adult emotional faces, may also be illuminating.

Additional limitations of the stimulus set include the exclusion of disgust and surprise picture sets. Surprise and disgust stimuli were not created because these emotions were not the focus of the research being conducted by the developers and so as not to overburden the child actors. In the future, creation of child disgust and surprise picture sets, pediatric picture sets with non-Caucasian children, and picture sets of younger children, including babies, would be useful for fu-

ture developmental neuroscience research. Using adult picture sets, researchers have begun to move beyond traditional facial expression picture sets and examine whether and how affective and cognitive neural pathways differ as parameters of the stimuli change. For example, researchers are looking at the impact of posed versus naturalistic pictures (McLellan *et al.*, [2009](#)), three-dimensional versus one-dimensional facial representations (Gur *et al.*, [2002](#)), static faces versus dynamic representations of emotions (Ambadar *et al.*, [2005](#); Schaefer *et al.*, [2009](#)) on emotion recognition and social processing. If use of the child static faces proves fruitful, development of child versions of these other types of stimuli would be warranted.

The NIMH-ChEFS picture set is a relatively large stimulus set with high quality, digital, color images of the emotional faces of children. The set includes neutral expressions and two gaze conditions (direct and averted). The set is freely available for scientists to download and use at no cost. The data from our validation study should give researchers confidence in their validity for use in affective and social neuroscience research.

Declaration of interest statement

Dr Egger receives grant funding from the National Institute of Mental Health (NIMH), National Institute on Drug Abuse (NIDA), Center for Disease Control (CDC), National Alliance for Research on Schizophrenia and Depression (NARSAD), and Autism Speaks. Dr Angold receives grant funding from the National Institute of Mental Health (NIMH) and National Institute on Drug Abuse (NIDA). Drs Pine, Nelson, Leibenluft, Ernst, and Towbin have no disclosures. The authors have no competing interests.

Supporting information

Supporting information may be found in the online version of this article.

Supporting information

[Click here for additional data file.](#)^(77K, xls)

Acknowledgements

This work was supported by the National Institute of Mental Health Grant # R01 MH081025. Tim Blitchington at the Center for Developmental Epidemiology developed the electronic rating program used in this study. The NIMH-ChEFS can be downloaded at <http://devepi.duhs.duke.edu/instruments.html>.

References

1. Adolphs R., Baron-Cohen S., Tranel D. (2002) Impaired recognition of social emotions following amygdala damage. *Journal of Cognitive Neuroscience*, 14(8), 1264–1274. [[PubMed](#)] [[Google Scholar](#)]
2. Adolphs R., Sears L., Piven J. (2001) Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience*, 13(2), 232–240. [[PubMed](#)] [[Google Scholar](#)]
3. Adolphs R., Tranel D., Damasio H., Damasio A. (1994) Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, 372(6507), 669–672. [[PubMed](#)] [[Google Scholar](#)]
4. 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(5), 403–410. [[PubMed](#)] [[Google Scholar](#)]
5. 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(1), 3–21. [[Google Scholar](#)]
6. Breiter H.C., Etcoff N.L., Whalen P.J., Kennedy W.A., Rauch S.L., 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(5), 875–887. [[PubMed](#)] [[Google Scholar](#)]
7. Brotman M.A., Guyer A.E., Lawson E.S., Horsey S.E., Rich B.A., Dickstein D.P., Pine D.S., Leibenluft E. (2008) Facial emotion labeling deficits in children and adolescents at risk for bipolar disorder. *The American Journal of Psychiatry*, 165(3), 385–389. [[PubMed](#)] [[Google Scholar](#)]
8. Brotman M.A., Rich B.A., Guyer A.E., Lunsford J.R., Horsey S.E., Reising M.M., Thomas L.A., Fromm S.J., Towbin K., Pine D.S., Leibenluft E. (2010) Amygdala activation during emotion processing of neutral faces in children with severe mood dysregulation versus ADHD or bipolar disorder. *The American Journal of Psychiatry*, 167(1), 61–69. [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
9. Calvo M.G., Lundqvist D. (2008) Facial expressions of emotion: Identification under different display-duration conditions. *Behavioral Research Methods*, 40(1), 109–115. [[PubMed](#)] [[Google Scholar](#)]
10. Calvo M.G., Nummenmaa L. (2009) Eye-movement assessment of the time course in facial expression recognition: Neurophysiological implications. *Cognitive, Affective, & Behavioral Neuroscience*, 9(4), 398–411. [[PubMed](#)] [[Google Scholar](#)]
11. Dalton K.M., Nacewicz B.M., Johnstone T., Schaefer H.S., Gernsbacher M.A., Goldsmith H.H., Alexander A.L., Davidson R.J. (2005) Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience*, 8(4), 519–526. [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
12. Ekman P. (1976) *Pictures of Facial Affect*, Palo Alto, CA: Consulting Psychologists Press. [[Google Scholar](#)]
13. Ekman P. (1994) Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2), 268–287. [[PubMed](#)] [[Google Scholar](#)]
14. Ekman P., Davidson R.J. (1994) *The Nature of Emotion: Fundamental Questions*, New York: Oxford University Press. [[Google Scholar](#)]
15. Ekman P., Friesen W.V. (1975) *Unmasking the Face*, Englewood Cliffs, NJ: Prentice Hall. [[Google Scholar](#)]

16. Evalynn J.M., Nigel W.B., David A.T.S., Peter F.L. (1996) Conditioning with facial expressions of emotion: Effects of cs sex and age. *Psychophysiology*, 33(4), 416–425. [[PubMed](#)] [[Google Scholar](#)]
17. Gao X., Maurer D. (2009) Influence of intensity on children's sensitivity to happy, sad, and fearful facial expressions. *Journal of Experimental Child Psychology*, 102(4), 503–521. [[PubMed](#)] [[Google Scholar](#)]
18. 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(2), 137–143. [[PubMed](#)] [[Google Scholar](#)]
19. Guyer A.E., McClure E.B., Adler A.D., Brotman M.A., Rich B.A., Kimes A.S., Pine D.S., Ernst M., Leibenluft E. (2007) Specificity of facial expression labeling deficits in childhood psychopathology. *Journal of Child Psychology and Psychiatry*, 48(9), 863–871. [[PubMed](#)] [[Google Scholar](#)]
20. Mazurski E.J., Bond N.W. (1993) A new series of slides depicting facial expressions of affect: A comparison with the pictures of facial affect series. *Australian Journal of Psychology*, 45(1), 41–47. [[Google Scholar](#)]
21. McLellan T., Johnston L., Dalrymple-Alford J., Porter R. (2009) Sensitivity to genuine versus posed emotion specified in facial displays. *Cognition & Emotion*. [[Google Scholar](#)]
22. Nelson E.E., McClure E.B., Monk C.S., Zarhn E., Leibenluft E., Pine D., Ernst M. (2003) Developmental differences in neuronal engagement during implicit encoding of emotional faces: An event-related fMRI study. *Journal of Child Psychology and Psychiatry*, 44(7), 1015–1024. [[PubMed](#)] [[Google Scholar](#)]
23. Palermo R., Coltheart M. (2004) Photographs of facial expression: Accuracy, response times, and ratings of intensity. *Behavioral Research Methods, Instruments & Computers*, 36(4), 634–638. [[PubMed](#)] [[Google Scholar](#)]
24. Pelphery K., Adolphs R., Morris J. (2004) Neuroanatomical substrates of social cognition dysfunction in autism. *Mental Retardation and Developmental Disabilities Research Reviews*, 10(4), 259–271. [[PubMed](#)] [[Google Scholar](#)]
25. Perlman S.B., Morris J.P., Vander Wyk B.C., Green S.R., Doyle J.L., Pelphrey K.A. (2009) Individual differences in personality predict how people look at faces. *PLoS One*, 4(6), e5952. [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
26. Schaefer A., Nils F., Sanchez X., Philippot P. (2009) Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition & Emotion*. [[Google Scholar](#)]
27. 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(4), 309–316. [[PubMed](#)] [[Google Scholar](#)]
28. Tottenham N., Tanaka J.W., Leon A.C., McCarry T., Nurse M., Hare T.A., Marcus D.J., Westerlund A., Casey B.J., Nelson C. (2009) The nimstim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, 168(3), 242–249. [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
29. Wagner H.L. (1993) On measuring performance in category judgment studies on nonverbal behavior. *Journal of Nonverbal Behavior*, 17(1), 3–28. [[Google Scholar](#)]
30. Wang L., Markham R. (1999) The development of a series of photographs of Chinese facial expressions of emotion. *Journal of Cross-Cultural Psychology*, 30(4), 397–410. [[Google Scholar](#)]