ijbd

International Journal of Behavioral Development
2021, Vol. 45(2) 146–152

DOI: 10.1177/0165025420935631
journals.sagepub.com/home/jbd


Validation of the child models of the Radboud Faces Database by children

Geraly Bijsterbosch,[1] Lynn Mobach,[1,2] Iris A. M. Verpaalen,[1] Gijsbert Bijlstra,[1] Jennifer L. Hudson,[2] Mike Rinck,[1] and Anke M. Klein[1]

Abstract

To draw valid and reliable conclusions from child studies involving facial expressions, well-controlled and validated (child) facial stimuli are necessary. The current study is the first to validate the facial emotional expressions of child models in school-aged children. In this study, we validated the Radboud Faces Database child models in a large sample of children ($N = 547$; 256 boys) aged between 8 and 12. In addition, associated validation measures such as valence, clarity, and model attractiveness were examined. Overall, the results indicated that children were able to accurately identify the emotional expressions on the child faces in approximately 70% of the cases. The highest accuracy rates were found for "happiness," whereas "contempt" received the lowest accuracy scores. Children confused the emotions "fear" and "surprise," and the emotions "contempt" and "neutral" with one another. Ratings of all facial stimuli are available (https://osf.io/7srgw/) and can be used to select appropriate stimuli to investigate the processing of children's facial emotional expressions.

Keywords

Facial emotions, child expressions, facial emotion recognition, Radboud Faces Database (RaFD), validation

Facial emotion recognition (FER) plays an important role in day-to-day social interactions: Facial expressions convey crucial information on the thoughts and feelings of an individual, which is necessary for successful communication between humans (Leppänen & Hietanen, 2001). The ability to differentiate between facial expressions develops between infancy and early adulthood (Lawrence et al., 2015). Child development researchers, therefore, highlight the importance of FER in the emotional development of children, and they often employ facial stimuli in their studies (e.g., Chronaki et al., 2014; Lawrence et al., 2015). However, almost all studies have included adult facial expressions when examining FER in children. This may make sense since children could be more competent in evaluating adult faces, based on their dependency on adults for their basics needs. However, solely including adult faces is problematic because of the repeated finding that children perceive and recognize emotional expressions from adults differently compared to emotional expressions from same-aged peers (e.g., Hills & Lewis, 2011; Rhodes & Anastasi, 2012). Moreover, from the age of six, most children will spend more time with peers outside their family (Rubin et al., 2011). During early adolescence, the importance of the peer group increases even further (Lansford et al., 2009). Therefore, using adult stimuli in child studies might lead to inaccurate representations of FER in children. Fortunately, there are a few studies that included child facial stimuli to study emotional development in children; however, these stimuli were mostly only validated by adults (e.g., Dalrymple et al., 2014; Marusak et al., 2013). To draw valid and reliable conclusions from studies involving facial expression paradigms in children, researchers need access to well-controlled and validated child facial stimuli that were validated by children. Therefore, the aim of the present study was to validate a child face database by children.

Currently, the authors are aware of the existence of five child face databases, including the NIMH Children Emotional Faces Picture Set (Egger et al., 2011), the Dartmouth Database of Children's Faces (Dalrymple et al., 2013), the Child Affective Facial Expression (CAFE) set (LoBue & Thrasher, 2015), the Child Emotion Picture Set (Romani-Sponchiado et al., 2015), and the Radboud Faces Database (RaFD; Langner et al., 2010). Of these databases, the only child face database that has been validated by both adults (e.g., LoBue & Trasher, 2015) and (preschool aged) children is the CAFE set (LoBue et al., 2018). They found that adults differ in recognition accuracy from preschoolers, with preschoolers still being able to recognize facial expressions above chance level. Thereby, this study suggested that the faces of the CAFE set could be applied to child research. However, this database was only validated in children aged between 3 years and 4 years, covering only a small target group in child research. This is problematic because child FER research is largely focused on school-aged children who are thought to have better FER abilities than preschool children (e.g., Chronaki et al., 2014). Therefore, the current study focuses on validating a child faces database, namely the RaFD, in 8- to 12-year-old children.

The RaFD (Langner et al., 2010) is a freely available database in which specific image characteristics are systematically varied,

[1] Radboud University, The Netherlands
[2] Macquarie University, Australia

Corresponding author:
Geraly Bijsterbosch, Department of Clinical Psychology, Behavioural Science Institute, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands.
Email: g.bijsterbosch@psych.ru.nl

including eight different facial expressions. It consists of an adult part, with Caucasian and Moroccan models, and a child part. The Radboud Faces Database child models (RaFD-C) consists of 1,200 pictures of 4 boys and 6 girls, all of them Caucasian (10 models × 8 emotions × 3 gaze directions × 5 camera angles). The models were trained to express the emotions "happiness," "sadness," "fear," "surprise," "disgust," "anger," "contempt," and a "neutral" facial expression, according to the Facial Action Coding System (Ekman et al., 2002, see Langner et al. (2010) for the procedure). The RaFD extends other databases, like FACES and the Karolinska Directed Emotional Faces database (KDEF), by including the emotion contempt. Although findings with regard to contempt are less clear and received less attention in research so far, this emotion continues to raise scientific interest in (child FER) research (e.g., Chaplin & Aldao, 2013; Fischer & Giner-Sorolla, 2016). Moreover, the pictures are highly standardized, in that models do not wear glasses or makeup, or have facial hair. Langner et al. (2010) concluded that the RaFD can be considered an adequate tool for research using facial stimuli. Specifically, adult participants could accurately identify the expressed emotions in 82% of the cases, for both child faces and adult faces. For the child faces, it was found that adult participants best recognized the emotion "happiness," followed by "surprise," whereas "contempt" was recognized worst. Additionally, recent research showed that school-aged children could accurately recognize 72% of the adult expressions (Verpaalen et al., 2019), but the child expressions have not been evaluated by children yet. The highly standardized features and qualities of the RaFD, together with its popularity (more than 1,100 citations since its availability; Rothermund & Koole, 2018) emphasize the value of choosing the RaFD-C as a nominee for validation by children as well.

The aim of the present study was to validate the RaFD-C in a large sample of school-aged children (aged 8–12). As the primary validation measure, we assessed the degree of agreement between the intended and chosen emotions. Overall, we expected the children to accurately recognize the expressed emotions, such that the agreement rate between the intended expressed emotion and the emotion chosen by the children would be above chance level. The chance level, for this specific choice task with nine (answer) options, is 11%. Hereafter, this agreement will be referred to as accuracy, to indicate how accurately children identified the displayed facial emotion expressions. Particularly, we expected that the emotion "happiness" would be recognized best compared to all other emotions, and "contempt" worst (Langner et al., 2010; Lawrence et al., 2015; LoBue et al., 2018). Finally, in order to assess the quality of the database and to provide researchers with overall ratings of the models, associated validation measures (model attractiveness, clarity, and valence of the expressed emotions) were examined (following Langner et al., 2010). Since clarity reflects whether an emotion is expressed well, and valence best represents the emotional experience of the emotion (Shuman, Sander, & Scherer, 2013), these stimulus features will support researchers in their stimulus selection.

Methods

Recruitment and Participants

A total of 547 children (256 boys) aged between 8 years and 12 years ($M = 9.9$, $SD = 1.2$) participated in the study. Children were recruited via nine elementary schools, representing both urban and rural regions of The Netherlands.[1] Schools were recruited via the personal networks of the involved authors and were selected based on their availability to participate in the study. The schools received an information letter about the study and received a phone call afterward to provide the principal and participating teachers the opportunity to ask questions about the study. If the school agreed to participate, an informed consent form was signed. The researchers supplied the schools with information letters and informed consent forms for parents and the children. Children were asked to indicate at the beginning of the session if they wanted to participate. Both parents and children of at least 12 years old were asked to sign active informed consent. The study was approved by the Ethics Committee of the Faculty of Social Sciences of the Radboud University, the Netherlands.

Materials

RaFD-C validation task. The RaFD-C (Langner et al., 2010) includes pictures of 10 different Caucasian (i.e., "white-colored skin") Dutch children (4 male and 6 female models) who vary in hair and eye color. Each child displays eight different emotions (anger, contempt, disgust, fear, happiness, neutrality, sadness, surprise; see also Langner et al., 2010). The included child models are aged between 7 years and 12 years, with the age of one child model being unknown. For the current study, 80 straight-gaze, frontal (90°) camera, pictures of the children were used. To not burden the participants too much, each child rated only a subset of the pictures. In total, 10 versions of a slideshow were created, containing a subset of the RaFD-C pictures with the constraint that each child model was included in four different slideshow versions. The first block of each slideshow started with 10 pictures of child models displaying a "neutral" expression. In order to rate the attractiveness of the models and to familiarize the children with the procedure, they were asked to indicate for each of the 10 "neutral" faces: "How attractive do you find this child?" (5-point Likert scale [1 = *not at all* to 5 = *very attractive*]).

During the second block, a subset of 32 pictures was presented, in which four different models expressed each of the eight emotions in a randomized order. The expression rating was forced-choice based. Children were instructed to answer three questions per picture: (1) "Which emotion does this person express?" (happiness, anger, fear, sadness, surprise, disgust, neutral, contempt, other); (2) "How clear do you find this emotion?" (5-point Likert scale [1 = *not at all* to 5 = *very clear*]); and (3) "How positive do you find this emotion?" (5-point Likert scale [1 = *not at all* to 5 = *very positive*]).

Procedure

The study was conducted at the participating schools during school hours in the children's regular classroom environment. The task and questionnaires were completed using pen and paper under close supervision. Before receiving instructions, the children were reminded about their right to stop at any time and/or ask questions. Researchers provided the children with the definitions of the emotions, as well as the specific questions and answering scales that were used. If necessary, the explanations were repeated, also during the tasks. Children were randomly assigned to one of the 10 slideshow versions, based on the date of testing. All children in one classroom received the same version. Children were seated so that they were

Table 1. Mean Hit Rates per Intended Emotion Category, Analyzed per Picture.

	Intended emotion								
	Happiness	Surprise	Anger	Neutral	Sadness	Disgust	Fear	Contempt	Range
Type of hit rate	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	Min–max
Raw hit rates (%)	92.34 (6)	64.36 (23)	77.31 (6)	80.54 (8)	75.35 (20)	68.34 (12)	62.89 (15)	39.30 (14)	0.00–100.00
Unbiased hit rates	0.88 (0.2)	0.56 (0.3)	0.64 (0.3)	0.64 (0.3)	0.62 (0.3)	0.56 (0.3)	0.49 (0.3)	0.32 (0.3)	0.00–1.00
Arcsin-transformed unbiased hit rates	1.29 (0.4)	0.67 (0.4)	0.79 (0.4)	0.81 (0.5)	0.78 (0.5)	0.68 (0.5)	0.57 (0.4)	0.38 (0.5)	0.00–1.57

Note. Mean hit rates are presented for all participants ($N = 547$). The variable range for all different intended emotions and different rates is included in the table and is the same for all emotions per type of hit rate.

able to clearly see the RaFD-C pictures, presented on a big screen in front of the class, and could not see each other's answers. The pictures were presented one by one, with four blocks of eight pictures. Each picture was presented until the last child finished all three corresponding questions in the paper booklet. The order of the choice options for the expressed emotions was kept constant.

The current study was part of a larger project on childhood anxiety. As a result, children also completed other questionnaires and tasks (see Baartmans et al., 2019; Klein et al., 2018). The completion of the RaFD-C validation task took approximately 30–45 min and the total testing time was approximately 60 min.

Data Preparation and Analyses

Calculations of accuracy rates, unbiased hit rates, and arcsin-transformed scores. First, raw hit rates were calculated separately for each emotion category. Raw hit rates are mean percentages of correct responses and indicate to which degree the chosen emotion is in agreement with the intended emotional expression. A formula developed by Wagner (1993) was used to create unbiased hit rates for tasks that use multiple choice answer formats in FER tasks (see also Elfenbein & Ambady, 2002). This formula corrects for possible answer habits, for example, if participants answer with one specific emotion category for all expressions. In line with Langner et al. (2010), the unbiased hit rates were calculated in two steps: First, a choice matrix was created with chosen emotion expressions as columns and intended emotional expressions as rows. Second, the number of ratings in each cell was squared and divided by the product of the corresponding row and column marginal values (Wagner, 1993). Finally, to correct for skewed variances of decimal fractions obtained from counts (Fernandez, 1992), unbiased hit rates were calculated and arcsin-transformed (recommended by Bromiley & Tacker, 2002; Bishara & Hittner, 2012; Winer, 1971). Arcsin-transformations are commonly used for proportional data (Bishara & Hittner, 2012), and they are a commonly applied "variance stabilizing transformers" for binominal distributions, especially with a large sample size (Bromiley & Tacker, 2002, p. 4).

Unusual response patterns. Next, the data were checked for unusual response patterns. Specifically, we checked if participants answered with a single emotion category, which cannot be corrected for by the used formula to correct for possible answer habits. None of the participants showed such an unusual response pattern.

Missing data. In total, 54 children did not report their age and/or gender. In the validation task, at the item level, 148 data points were missing (0.8% of the data points), 127 (0.7%) clarity ratings were missing, as were 190 (1.1%) data points in the positivity ratings. Only one subject was missing entirely. The answers of the participants were scored as correct (1) or incorrect (0). In total, each participant rated 80 pictures; every emotion (expressed by a different model) was shown 4 times. If the participant rated all four expressions of a specific emotion correctly, the unbiased score was 1.00 (4/4; based on the formula developed described in the data preparation section and recommended by Bromiley & Tacker, 2002; Bishara & Hittner, 2012; Wagner, 1993; Winer, 1971). If there was a missing value, and the participant, therefore, rated a specific emotion only 3 times, this was taken into account while computing the unbiased score for that emotion. In this case, the maximum score for the participant was 0.75 (if all three were correct), that is, the missing data point was treated as an incorrect response. The data used for the final analyses (i.e., arcsine transformed scores) did not have missing data points.

Analyses. To examine the overall accuracy across the different emotions, repeated-measures analyses of variance (ANOVAs) were computed in SPSS 21 (IBM Corp, 2012). Separate ANOVAs were computed for clarity and valence with emotion (happiness, surprise, neutral, sadness, anger, disgust, fear, and contempt) as a within-subject factor. Intraclass correlation coefficients (ICCs) were calculated to determine the interrater reliability between the children's accuracy rates on valence, clarity, and model attractiveness.

Results

Accuracy Rates of the Expressed Emotions

Table 1 presents an overview of the means and standard deviations of all accuracy measures (i.e., raw hit rates, unbiased hit rates, arcsin-transformed hit rates) separately for each expression.[2] For an overview of all choice rates and significant differences from chance level, see Table 2. To assess accuracy, a repeated-measures ANOVA was computed on the arcsin-transformed unbiased hit rates with emotion (happiness, surprise, neutral, sadness, anger, disgust, fear, and contempt) as a within-subject factor. As the assumption of sphericity was violated, $\chi^2(27, N = 547) = 186.62$, $p < .001$, the Huynh–Feldt correction was applied.[3] The repeated-measures ANOVA yielded a significant within-subject effect of emotion, $F(6.40, 3{,}496.23) = 234.10$, $p < .001$, $\eta_p^2 = .30$, suggesting that particular emotions were significantly better recognized than others.[4] To test which specific emotions were better recognized than others, we calculated associated deviation

Table 2. Choice Percentages per Intended and Chosen Emotion Category.

	Chosen emotion								
Intended emotion	Happiness	Anger	Neutral	Surprise	Sadness	Disgust	Fear	Contempt	Other
Happiness	92.34	0.42	2.54	0.59	0.09	0.72	0.57	1.37	1.36
Surprise	0.48	77.31	1.53	0.46	0.20	0.94	14.79	0.89	3.40
Neutral	1.00	3.28	80.54	1.03	1.75	1.86	1.11	3.34	6.23
Sadness	0.31	6.14	2.53	64.36	2.35	7.16	9.64	3.49	4.13
Anger	0.31	5.65	1.29	1.70	75.35	4.42	1.83	4.58	4.70
Disgust	0.43	6.03	1.11	0.68	11.53	68.35	0.77	5.73	5.47
Fear	1.01	16.66*	0.60	1.36	4.84	7.26	62.89	1.07	4.29
Contempt	2.73	8.29	22.48**	1.08	0.92	3.93	0.96	39.30	20.17**

Note. Choice percentages are presented for all participants ($N = 547$). The cells contain raw hit percentages, hits are marked in light gray. Contoured cells, noted with asterisks, contain percentages of emotion confusions above chance level (*$p < .05$, **$p < .01$, ***$p < .001$).

contrasts from the grand mean accuracy rate ($M_{grand} = .75$). Taking the high number of ratings per emotion into account, only contrasts with an effect size of $\eta_p^2 \geq .10$ are reported (in line with Langner et al., 2010). "Happiness" was better recognized ($M = 1.29$) than the grand mean accuracy rate between all intended and chosen emotions $F(1,546) = 1{,}258.73$, $p < .001$, $\eta_p^2 = .70$. On the contrary, "contempt" was significantly worse recognized ($M = .38$) than the grand mean accuracy rate, $F(1,546) = 73.24$, $p = .482$, $\eta_p^2 = .48$.

To compare the different emotions with each other, we computed Bonferroni-corrected pairwise comparisons. These comparisons indicated that "happiness" was significantly better recognized than all other emotions ($p < .001$). Next, the emotions "neutral," "sadness," and "anger" received the second-highest overall accuracy rates and differed significantly from the emotions with lower accuracy rates ($p < .001$). Accuracy for "disgust" and "surprise" was significantly higher than for "fear" and "contempt" ($ps < .001$). Finally, "fear" was found to have a significantly higher accuracy rate than "contempt" ($p < .001$).

Interestingly, the results indicated a pattern of "emotion confusions" for specific emotions, which means that children often misidentified some of the intended emotions with a specific other emotion above the chance level of 11% (in this forced-choice task with nine answer options). The rationale for these patterns was that it might be possible that participants may choose particular emotions more frequently than other emotions when they are not sure of their response. One-sample *t*-tests showed three significant emotion confusions: Children significantly misidentified the indented expression of "fear" as "surprise" ($p = .023$), and they significantly misidentified "contempt" as "neutral" ($p = .008$) or "other" ($p = .002$).

Associated Validation Measures

Clarity and valence ratings of the pictures. Next, mean clarity and valence ratings were computed (see Table 3). The ANOVA for clarity revealed a significant main effect of emotion for clarity ratings, $F(7,72) = 13.94$, $p < .001$, $\eta_p^2 = .58$. This indicates that some emotions were rated as significantly clearer than other emotions. Bonferroni-corrected pairwise comparisons revealed that "happiness" was rated as significantly more clear than all other emotions (all $ps < .01$) and "contempt" was rated as less clear compared to all other emotions (all $ps < .001$). Moreover, "sadness" was rated as significantly less clear as "surprised" ($p = .046$). All other emotions did not differ significantly from each other. The second ANOVA, with regard to the emotional valence of the pictures, revealed that some emotions were rated as significantly more positive than other emotions, $F(7,72) = 177.17$, $p < .001$, $\eta_p^2 = .95$. Bonferroni-corrected pairwise comparisons showed that "happiness" was rated as significantly more positive than all other emotions (all $ps < .001$). Moreover, "neutral" and "surprise" were rated as significantly more positive than "contempt," "fear," "disgust," "sadness," and "anger" (for both "neutral" and "surprise" all $ps < .001$). Next, the valence rates for "fear" were significantly higher than for "sadness" ($p = .001$) and "anger" ($p < .001$), but not for "disgust." However, "disgust" was found to have a significantly higher valence rate than "anger" ($p < .001$).

Table 3. Mean Clarity and Valence Ratings of the Pictures.

	Clarity		Valence	
Intended emotion	*M* (*SD*)	Min–max	*M* (*SD*)	Min–max
Happiness	4.14 (0.5)	3.25–4.50	4.05 (0.2)	3.61–4.27
Sadness	3.38 (0.3)	2.91–3.78	2.37 (0.1)	2.23–2.68
Surprise	3.69 (0.1)	3.57–3.85	2.94 (0.1)	2.77–3.11
Neutral	3.41 (0.3)	2.85–3.85	3.03 (0.1)	2.79–3.19
Anger	3.42 (0.5)	2.52–3.91	2.26 (0.2)	2.10–2.52
Disgust	3.39 (0.2)	3.08–3.81	2.47 (0.0)	2.36–2.52
Fear	3.54 (0.4)	2.75–3.89	2.57 (0.1)	2.45–2.65
Contempt	2.72 (0.2)	2.38–2.99	2.71 (0.1)	2.46–3.00

Note. Mean clarity and valence ratings are presented from all participants ($N = 547$) for the different pictures ($N = 80$). Each mean and standard deviation, for the separate emotions, is based on clarity and the valence ratings of 10 pictures ($n = 10$). The variables ranged between 1 and 5.

Interrater reliability. To objectively select the appropriate ICCs, the model fit (deviance information criterion) was calculated for all possible ICC models. The model with the best fit to our data was the ICC 2, applying a two-way cross-classified multilevel model.[5] The corresponding ICCs (2, *x*) for valence, clarity, and model attractiveness are reported in Table 4. Generally, the calculated ICCs confirm the expected high agreement of the raters for clarity, valence, and model attractiveness.

Discussion

The goal of the present study was to validate the RaFD-C by children. As the primary validation measure, we investigated to which

Table 4. ICCs for Clarity and Valence of the Emotions, and Model Attractiveness.

Dimension	ICC (2,1)	ICC (2, $k = 547$)	ICC (2, k adjusted for missings)
Clarity	0.14	0.99	0.97
Valence	0.31	0.96	0.99
Attractiveness	0.01	0.85	0.85

Note. ICCs are presented for all participants ($N = 547$). ICC (2, 1) = intraclass correlation coefficient, each subject is measured by each rater with reliability estimated from a single measurement, ICC (2, k) = as before, but here reliability is calculated by imputing the average of the k raters' measurements.

degree children aged 8–12 years labeled the emotions in agreement with the intended emotions. Finally, clarity and valence of the emotions, and model attractiveness were investigated. This study establishes childhood norms for the recognition of seven emotions and a neutral expression using the RaFD-C.

In the current study, we observed that children were able to correctly identify, on average, 70% of the emotions. Overall, this means that the children were relatively accurate in recognizing the emotions from the child models. The agreement rate in adults rating the RaFD-C was approximately 82% (Langner et al., 2010), and 72% in children who rated pictures of RaFD adult expressions (Verpaalen et al., 2019). In line with our hypotheses and consistent with other studies, we found that "happiness" was recognized best, whereas "contempt" was least accurately identified. Specific differences between emotions were found in the following rank order: "Happiness" accuracy rates were higher than all other emotions, followed by "neutral," "sadness," and "anger," next by "disgust" and "surprise," then by "fear," and finally "contempt" accuracy rates were lower than all others.

These results roughly follow the pattern found in the RaFD validation study by adults (Langner et al., 2010), suggesting continuity in the development of FER. They also concur with the results of Verpaalen et al. (2019) and the FER-study in 6- to 15-year-olds by Lawrence et al. (2015), who investigated the development of emotion recognition in adult faces through childhood and adolescence. Also, the child faces validation study in preschool children by LoBue et al. (2018) approximately found the same emotion accuracy patterns as in the current study. For a comparison of the current results with the agreement rates of the validation study in adults (Langner et al., 2010) and the validation study in preschool children (LoBue et al., 2018), see Table 5. Moreover, our results replicate previous findings that "happiness" is easiest to recognize for children (e.g., Lawrence et al., 2015). The finding that the emotions "surprise," "disgust," and "fear" were found to be disproportionally difficult to recognize resonates with earlier studies as well (Gagnon et al., 2014; Rozin et al., 2005). It is likely that these complex emotions are generally more difficult to interpret, more easily confused with other emotions, and mature significantly through the course of late childhood into adolescence (Gagnon et al., 2010; Lawrence et al., 2015). The similarity with other research using child participants suggests that the observed differences between the expressions are not likely to be a specific characteristic of the RaFD-C.

The current study had several strengths. First, this study is the first of its kind to have a child faces database validated by school-aged children (8–12 years), including an assessment of all six basic emotions, plus the emotion "contempt" and a "neutral" expression.

Table 5. Raw Hit Rates for the RaFD-C Child and Adult Validations, and the CAFE Database Validated by Children.

	RaFD		CAFE set
Emotion category	Child validation (current study) *M* (*SD*)	Adult validation (Langner et al., 2010) *M* (*SD*)	Child validation (LoBue et al., 2018) *M* (*SD*)
Happiness	92 (6)	97 (6)	78 (42)
Surprise	64 (23)	91 (25)	61 (49)
Anger	77 (6)	89 (8)	65 (48)
Neutral	81 (8)	84 (16)	44 (50)
Sadness	75 (20)	75 (14)	49 (50)
Disgust	68 (12)	83 (18)	47 (50)
Fear	63 (15)	79 (13)	26 (44)
Contempt	40 (14)	59 (19)	NA

Note. Raw hit rates for the current validation study are presented for all participants ($N = 547$) in the column on the left, and the raw hit rates for the adult validation study by Langner et al. (2010) in the middle ($N = 276$), whereas the raw hit rates for the child validation study by LoBue et al. (2018) is presented in the column on the right ($N = 58$). The variable range for the raw hit rates for all three studies is between 0 and 100. The included agreement rates are the raw hit rates and standard deviations per emotion, the agreement rates from the child validation of the CAFE set are the scores from the full set at Time 1. RaFD = Radboud Faces Database; CAFE = the Child Affective Facial Expression Set.

This provides researchers with a validated child faces database, allowing them to draw valid and reliable conclusions from studies involving facial expression paradigms in children. The current findings can be generalized to other same-aged Caucasian child faces. However, limitations of the current study need to be mentioned as well. First, conducting the current study in a class-wise manner implied some practical constraints. Fast-responding children had to wait before the next picture was presented. This might have led to disinterest and consequently, interrupted attention patterns in these children (Malkovsky et al., 2012). Besides, children who needed more time to answer the questions might have felt pressured or hurried because other children had to wait for them to finish. Both situations might have negatively biased the FER accuracy rates of the current study; either lacking attention or time pressure may have led children to select other answers than they would have chosen if they had followed their own pace. Using a forced-choice task could possibly rely on the development of executive functioning, meaning that a process of elimination could have been used to deduce which emotion was presented. A review on developmental studies regarding deductive reasoning stated that that deductive reasoning is a characteristic of the concrete operational period, between 7 years and 11 years (Jansson, 1974). However, unbiased hit rates and the option *other* in the answering format were used to minimize this bias possibility. Adding *other* to the answering format counteracts the bias in forced-choice FER paradigms (Frank & Stennet, 2001). Next to this, it could be that the used method to handle the missing data inherently also has led to biased results. However, the number of missing data points is very small, so is unlikely to have led to biased estimates, and we are being conservative using this method. Finally, it would have been valuable to collect more detailed information about the race/ethnicity and social economic status of our participants to be able to explore possible differences in FER between subgroups of our sample. Moreover, it was not registered why children did not participate in the current study. This would be valuable to determine the representativeness of the sample but also to control for possible

confounds related to participation refusal. The generalizability of the results is an important point to be addressed in future research because we currently cannot generalize to emotion recognition in non-Caucasian models. It would be beneficial for future studies to validate a child faces database representing a variety of ethnicities, together with an ethnically diverse sample of participants.

The validation data per picture are provided freely online (see https://osf.io/7srgw/), enabling researchers to include specific pictures as stimuli and select the most appropriate items for their research. Based on this study, some recommendations about selecting stimuli can be made. Specifically, we recommend to be cautious in combining (pictures with) the emotions "fear" and "surprise," and the emotions "contempt" and "neutral" together in one study since these were found to be confused with each other. Finally, researchers who are interested in pictures that are easily recognizable and do not include ambiguity of the different emotional expressions can use the available validation data to select those pictures that were identified with at least 60% accuracy by the children. Although this will lead to a more limited choice in terms of suitable pictures, this subset of 60 pictures will still enable researchers to select pictures based on criteria of their choice.

In conclusion, the current study indicated that 8- to 12-year-old children were able to accurately identify approximately 70% of the RaFD-C pictures. The differences in the child FER rates compared to the FER rates of adults emphasize the importance of a validation study of child faces stimuli from a child's perspective. Additionally, the results from the current study indicate that subtle emotion differences should be considered when studying FER in children, as reflected in our recommendations above. As this is the first study that validated the RaFD-C in children, we expect that researchers focusing on FER in childhood will benefit from this study for further and better understanding of emotion recognition in children.

Acknowledgments

We are grateful to the schools and the children for their participation in the study. We thank Pierre Souren for calculating the inter-rater reliability indices and Giovanni ten Brink for his help with the data preparation. One of the authors was supported by a Niels Stensen Fellowship.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Geraly Bijsterbosch https://orcid.org/0000-0003-4390-9794
Gijsbert Bijlstra https://orcid.org/0000-0002-0827-7376

Notes

1. Student populations of Dutch nonprivate elementary schools have mixed socioeconomic characteristics and a Caucasian ethnic majority (Centraal Bureau voor de Statistiek, 2019). Therefore, we can assume that the included schools in the current study represented varied socioeconomic characteristics similar to other Dutch elementary schools. However, not all children from a class participated in the study. Generally, the reason for refusal to participate was decided by the researcher on site as this was mostly due to parents forgetting to hand in the informed consent form within due time. Overall, approximately 121 (22.2%) of the children were in the third grade, 145 (26.7%) in the fourth grade, 128 (23.5%) in the fifth grade, and 150 (27.6%) in the sixth grade.
2. Due to the importance of age for facial emotion recognition (FER) performance (Field & Lester, 2010; Martin & Ruble, 2013) and given that researchers may want to select the best fitting stimuli for different age groups, the current study also investigated age group differences in FER performance. For an overview of the means and standard deviations of the accuracy for each emotion, separately for age, see Table S1 in the online supplementary materials (https://osf.io/7srgw/).
3. If the data violate the sphericity assumption, corrections should be made to produce a valid F-ratio. The used correction is based on the estimates of sphericity (ϵ). When $\epsilon > .75$, it is advised to use the Huynh–Feldt correction (Field, 1998; Girden, 1992). In the current study, it was found that $\epsilon = .915$, resulting in the valid usage of the Huynh–Feldt correction to correct for sphericity violations.
4. As an additional check to justify the use of arcsin-transformed scores in the current study, the exact same repeated-measures ANOVA was conducted with the raw hit rates as dependent variable. This analysis yielded similar results.
5. Cross-classified multilevel Markov Chain Monte Carlo intraclass correlation coefficients (ICCs) were calculated. This type of ICC is uncommon, however, it was decided to use this type in this study since this type fitted the data best.

References

Baartmans, J. M. D., Rinck, M., Hudson, J. L., Lansu, T. A. M., van Niekerk, R. E., Bögels, S. M., & Klein, A. M. (2019). Are socially anxious children really less liked, or do they only think so? *Cognitive Therapy and Research*, *43*, 1043–1050. https://doi.org/10.1007/s10608-019-10028-9

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, *17*(3), 399–417. https://doi.org/10.1037/a0028087

Bromiley, P. A., & Thacker, N. A. (2002). *The effects of an arcsin square root transform on a binomial distributed quantity* (TINA memo no. 2002-007). University of Manchester.

Centraal Bureau voor de Statistiek. (2019). *Leerlingen in (speciaal)basisonderwijs; migratieachtergrond, woonregio*. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83295NED/table?ts=1572185466637

Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: A meta-analytic review. *Psychological Bulletin*, *139*(4), 735–765. https://doi.org/10.1037/a0030737

Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., & Sonuga--Barke, E. J. (2014). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *British Journal of Developmental Psychology*, *33*(2), 218–236. https://doi.org/10.1111/bjdp.12075

Dalrymple, K. A., Gomez, J., & Duchaine, B. (2013). The Dartmouth database of children's faces: Acquisition and validation of a new face stimulus set. *PLoS ONE*, *8*(11), e79131. https://doi.org/10.1371/journal.pone.0079131

Dalrymple, K., Garrido, L., & Duchaine, B. (2014). A dissociation between face perception and face memory in adults, but not children, with developmental prosopagnosia. *Journal of Vision*, *14*(10), 1434–1434. https://doi.org/10.1167/14.10.1434

Egger, H. L., Pine, D. S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K. E., & Angold, A. (2011). The NIMH child emotional faces picture set (NIMH-ChEFS): A new set of children's facial emotion stimuli. *International Journal of Methods in Psychiatric Research*, *20*(3), 145–156. https://doi.org/10.1002/mpr.343

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system: The manual*. Research Nexus.

Elfenbein, H. A., & Ambady, N. (2002). Is there an in-group advantage in emotion recognition? *Psychological Bulletin*, *128*(2), 243–249. https://doi.org/10.1037/0033-2909.128.2.243

Fernandez, G. C. (1992). Residual analysis and data transformations: Important tools in statistical analysis. *HortScience*, *27*(4), 297–300.

Field, A. P. (1998). A bluffer's guide to . . . sphericity. *British Psychological Society-MSC Newsletter*, *6*(1), 13–24. http://www.discoveringstatistics.com/

Field, A. P., & Lester, K. J. (2010). Is there room for 'development' in developmental models of information processing biases to threat in children and adolescents? *Clinical Child and Family Psychology Review*, *13*(4), 315–332. https://doi.org/10.1007/s10567-010-0078-8

Fischer, A., & Giner-Sorolla, R. (2016). Contempt: Derogating others while keeping calm. *Emotion Review*, *8*(4), 346–357. https://doi.org/10.1177/1754073915610439

Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology*, *80*(1), 75–85. https://doi.org/10.1037/0022-3514.80.1.75

Gagnon, M., Gosselin, P., Hudon-ven der Buhs, I., Larocque, K., & Milliard, K. (2010). Children's recognition and discrimination of fear and disgust facial expressions. *Journal of Nonverbal Behavior*, *34*(1), 27–42. https://doi.org/10.1007/s10919-009-0076-z

Gagnon, M., Gosselin, P., & Maassarani, R. (2014). Children's ability to recognize emotions from partial and complete facial expressions. *The Journal of Genetic Psychology*, *175*(5), 416–430. https://doi.org/10.1080/00221325.2014.941322

Girden, E. R. (1992). *ANOVA: Repeated measures* (pp. 7–84). Sage University Press Series on Quantitative Applications in the Social Sciences.

Hills, P. J., & Lewis, M. B. (2011). Rapid communication: The own-age face recognition bias in children and adults. *Quarterly Journal of Experimental Psychology*, *64*(1), 17–23. https://doi.org/10.1080/17470218.2010.537926

IBM Corp. (2012). *IBM Corp statistics for windows* (Version 21.0).

Jansson, L. C. (1974). *The development of deductive reasoning: A review of the literature*. Preliminary version.

Klein, A. M., Bakens, R., van Niekerk, R. E., Ouwens, M. A., Rapee, R. M., B, S. M., ö, E. S., gels, M., Becker, O., & Rinck, R. (2018). Children with generalized anxiety disorder symptoms show a content-specific interpretation bias using auditory stimuli. *Journal of Behavior Therapy and Experimental Psychiatry*, *61*, 121–127. https://doi.org/10.1016/j.jbtep.2018.06.011

Langner, G., Dotsch, D. H. J., Bijlstra, S. T., Wigboldus, A., Hawk, J. E., & van Knippenberg, L. A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, *24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Lansford, S., Killeya-Jones, P. R., Miller, K., & Costanzo, R. (2009). Early adolescents' social standing in peer groups: Behavioral correlates of stability and change. *Journal of Youth and Adolescence*, *38*(8), 1084–1095. https://doi.org/10.1007/s10964-009-9410-3

Lawrence, D., Campbell, J. M., & Skuse, J. K. (2015). Age, gender, and puberty influence the development of facial emotion recognition. *Frontiers in Psychology*, *6*(761). https://doi.org/10.3389/fpsyg.2015.00761

Leppänen, J. M., & Hietanen, M. (2001). Emotion recognition and social adjustment in school–aged girls and boys. *Scandinavian Journal of Psychology*, *42*(5), 429–435. https://doi.org/10.1111/1467-9450.00255

LoBue, C., Baker, V., & Thrasher, C. (2018). Through the eyes of a child: Preschoolers' identification of emotional expressions from the child affective facial expression (CAFE) set. *Cognition and Emotion*, *32*(5), 1122–1130. https://doi.org/10.1080/02699931.2017.1365046

LoBue, E., & Thrasher, C. (2015). The child affective facial expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, *5*, 1532. https://doi.org/10.3389/fpsyg.2014.01532

Malkovsky, Y., Merrifield, J., Goldberg, C. L., & Danckert, D. N. (2012). Exploring the relationship between boredom and sustained attention. *Experimental Brain Research*, *221*(1), 59–67. https://doi.org/10.1007/s00221-012-3147-z

Martin, H. A., & Ruble, J. M. (2013). Patterns of gender development. *Annual Review of Psychology*, *61*, 353–381. https://doi.org/10.1146/annurev.psych.093008.100511

Marusak, M. E., Carré, M. G., & Thomason, J. S. (2013). The stimuli drive the response: An fMRI study of youth processing adult or child emotional face stimuli. *NeuroImage*, *83*, 679–689. https://doi.org/10.1016/j.neuroimage.2013.07.002

Rhodes, A., & Anastasi, B. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174. https://doi.org/10.1037/a0025750

Romani-Sponchiado, C., Sanvicente-Vieira, D., Mottin, A., Hertzog-Fonini, K., & Arteche, S. L. (2015). Child emotions picture set (CEPS): Development of a database of children's emotional expressions. *Psychology & Neuroscience*, *8*(4), 467–478. https://doi.org/10.1037/h0101430

Rothermund, P., & Koole, C. (2018). Three decades of cognition & emotion: A brief review of past highlights and future prospects. *Cognition and Emotion*, *32*(1), 1–12. https://doi.org/10.1080/02699931.2018.1418197

Rozin, L., Taylor, G., Ross, A., Bennett, K. H., & Hejmadi, W. M. (2005). General and specific abilities to recognise negative emotions, especially disgust, as portrayed in the face and the body. *Cognition & Emotion*, *19*(3), 397–412. https://doi.org/10.1080/02699930441000166

Rubin, B., Bukowski, I. A. M., & Laursen, G. (2011). *Handbook of peer interactions, relationships, and groups* (1st ed.). Guilford Press.

Shuman, V., Sander, D., & Scherer, K. L. (2013). Levels of valence. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00261

Verpaalen, L., Bijsterbosch, G., Mobach, M., Bijlstra, A., Rinck, H. L., & Klein, B. (2019). Validating the Radboud Faces Database from a child's perspective. *Cognition and Emotion*. https://doi.org/10.1080/02699931.2019.1577220

Wagner (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3–28. https://doi.org/10.1007/BF00987006

Winer (1971). *Statistical principles in experimental design* (2nd ed.). McGraw-Hill.