UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Validating the Radboud faces database from a child's perspective

Verpaalen, I.A.M.; Bijsterbosch, G.; Mobach, L.; Bijlstra, G.; Rinck, M.; Klein, A.M.

DOI
10.1080/02699931.2019.1577220

Publication date
2019

Document Version
Final published version

Published in
Cognition and Emotion

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Verpaalen, I. A. M., Bijsterbosch, G., Mobach, L., Bijlstra, G., Rinck, M., & Klein, A. M. (2019). Validating the Radboud faces database from a child's perspective. *Cognition and Emotion*, *33*(8), 1531-1547. https://doi.org/10.1080/02699931.2019.1577220

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:15 Sep 2026

COGNITION AND EMOTION
2019, VOL. 33, NO. 8, 1531–1547
https://doi.org/10.1080/02699931.2019.1577220

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Validating the Radboud faces database from a child's perspective

Iris A. M. Verpaalen [a], Geraly Bijsterbosch[a], Lynn Mobach [a], Gijsbert Bijlstra[a], Mike Rinck[a] and Anke M. Klein[a,b]

[a]Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands; [b]Developmental Psychology, Universiteit van Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Facial expressions play a central role in diverse areas of psychology. However, facial stimuli are often only validated by adults, and there are no face databases validated by school-aged children. Validation by children is important because children still develop emotion recognition skills and may have different perceptions than adults. Therefore, in this study, we validated the adult Caucasian faces of the Radboud Faces Database (RaFD) in 8- to 12-year-old children ($N = 652$). Additionally, children rated valence, clarity, and model attractiveness. Emotion recognition rates were relatively high (72%; compared to 82% in the original validation by adults). Recognition accuracy was highest for happiness, below average for fear and disgust, and lowest for contempt. Children showed roughly the same emotion recognition pattern as adults, but were less accurate in distinguishing similar emotions. As expected, in general, 10- to 12-year-old children had a higher emotion recognition accuracy than 8- and 9-year-olds. Overall, girls slightly outperformed boys. More nuanced differences in these gender and age effects on recognition rates were visible per emotion. The current study provides researchers with recommendation on how to use the RaFD adult pictures in child studies. Researchers can select appropriate stimuli for their research using the online available validation data.

ARTICLE HISTORY
Received 29 June 2018
Revised 22 January 2019
Accepted 27 January 2019

KEYWORDS
Emotion; child development; facial emotion recognition; Radboud faces database (RaFD); validation

Facial Emotion Recognition (FER) is crucial for children's social functioning in numerous daily life domains (Leppänen, 2011). Understanding the emotions communicated by facial expressions is needed for information processing, interpersonal communication, memory, and empathy (Calder & Young, 2005; Grossmann & Johnson, 2007). The importance of FER has led to extensive use of emotional facial stimuli in child (developmental) studies. Studies in this line of research mostly utilise facial stimuli validated by adults, because, to the best of our knowledge, there are no published studies that had children validate adult face databases. However, basing emotion research in child samples on adult validation data has consequences for the validity and reliability of the results of these studies, as research shows that children process facial expressions differently and more slowly than adults (Batty & Taylor, 2006; De Sonneville et al., 2002). To improve child FER research, a face database validated by children is needed. Validating children's responses to adults' expressions is important because the emotions that adults display, especially adult caregivers, are of major importance for children's emotional development. Moreover, children cannot rely on the relative advantage of the "own-age bias" when processing adult faces (Rhodes & Anastasi, 2012), the finding that emotions on faces of age-peers are easier to decode. The current study is the first to validate adult facial emotion stimuli in school-aged children.

Numerous databases with emotional faces of adults exist, such as the Ekman-Friesen Pictures of

CONTACT Iris A. M. Verpaalen i.verpaalen@psych.ru.nl

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Facial Affect (Ekman & Friesen, 1976), KDEF (Goeleven, De Raedt, Leyman, & Verschuere, 2008), FACES (Ebner, Riediger, & Lindenberger, 2010), ADFES (Van der Schalk, Hawk, Fischer, & Doosje, 2011), and the RaFD (Langner et al., 2010). The RaFD is a high-quality picture set of 67 models, including Caucasian men, women and children, and Moroccan-Dutch men. Besides including the six basic emotions (anger, disgust, fear, happiness, sadness, surprise,) and a neutral expression, based on prototypes of the facial action coding system (FACS; Ekman, Friesen, & Hager, 2002), the database also includes contempt. As contempt is proposed to be an important emotion that evokes a strong, ongoing scientific interest, including contemptuous faces extends the range of emotions in comparison to the FACES and KDEF. (e.g. Gervais & Fessler, 2017). Like disgust, contempt is thought to be a "hostile emotion", however, underpinned by different cognitive-motivational components and social function (for a discussion on features of contempt see, Miceli & Castelfranchi, 2018). The RaFD is highly standardised (no glasses, make-up or facial hair), includes models with modern haircuts and neutral clothing compared to the Ekman-Friesen Pictures of Facial Affect, and takes eye-gaze and camera position into account. These qualities of the RaFD, as well as its frequent use (over 1000 citations since its presentation) make the RaFD a good candidate for further validation research.

Even though RaFD stimuli are often used in child samples (e.g. Brown et al., 2014; Corbett et al., 2016), so far, the RaFD has only been validated by adults. Langner et al. (2010) found that adults accurately categorised the expressed emotions in 82% of the cases. Additionally, they showed that adults were significantly better in recognising happiness (98%) and significantly worse in recognising contempt (50%), compared to other emotions. In comparison, this ranged between 43% (fear) and 93% (happiness) with an average accuracy of 72% for the KDEF (Goeleven et al., 2008), between 68% (disgust) and 96% (happiness) for the FACES database (Ebner et al., 2010), and between 68% (contempt) and 91% (happiness) with an average accuracy of 81% for the ADFES (Van der Schalk et al., 2011). Furthermore, all RaFD emotions were perceived to be fairly clearly expressed, although contempt was perceived as somewhat less clear. Also, valence ratings of the emotions corresponded to intuitive expectations; happiness was rated as the only clearly positive expression, neutral and surprise were rated as fairly neutral, and the other emotions were rated as negative (see supplementary online materials for exact ratings). As shown by a high interrater reliability, adult participants agreed on the valence and clarity of the pictures and attractiveness of the models. Langner et al. (2010) concluded that the RaFD has equally high, or higher, recognition rates compared to other databases and is a useful tool for different fields of research using facial stimuli. However, whether children recognise the emotional stimuli as accurately as adults is unknown. Without knowledge about validity in children, researchers may select suboptimal stimuli or fail to measure what they want to investigate.

Recently, some attention has been drawn to this issue by LoBue, Baker, and Thrasher (2018), who validated child pictures in pre-school children. Their sample consisted of 3- to 4-year old children and investigated the same emotions as the RaFD contains, except for contempt. In their study, children verbally labelled the emotions, and if they could not label them, they were asked if the emotion matched one of the provided emoticons. The results from children who only used verbal labelling and children who also used matching did not differ significantly, the overall recognition rate was fairly low (54%). This indicates the need to test whether a standard faces database is suitable to use for school-aged children, who are more developed and may therefore perform better on a verbal labelling task. Moreover, many studies focus on recognising adult, instead of child, facial expressions (e.g. Durand, Gallay, Seigneuric, Robichon, & Baudouin, 2007; Gao & Maurer, 2009; Lawrence, Campbell, & Skuse, 2015). Therefore, the current study tests whether children can also accurately label the facial expressions of adults.

Children do not form a homogenous group when it comes to FER; developmental factors such as age and gender seem to play an important role. It should be noted however, that the knowledge about FER in children is based on research with stimuli that were only validated by adults, and results should be interpreted with care. Regarding age effects, several empirical studies showed that FER accuracy increases over the course of late childhood, but with distinct developmental pathways for different emotions (Mancini, Agnoli, Baldaro, Bitti, & Surcinelli, 2013). This development can be explained by a differentiation account of emotion categories (Bullock & Russell, 1984). According to this account, FER becomes increasingly specific with age; children first distinguish positive from negative valence, then low from high arousal,

and later they learn more specific categorisations. In line with these results, recognition of happiness and sadness seems to develop before fear and disgust (Durand et al., 2007). Specifically, Durand et al. (2007) found that 5-year-old children could recognise happy and sad facial pictures in around 90% of the cases, whereas these accuracy levels for anger and fear were reached around the age of 9. Disgust was not recognised at this high level until children were 11 years old. However, other studies did not find an improvement in late childhood, possibly due to (1) the used (complicated) tasks that involved matching a facial expression to descriptions of emotions (Gagnon, Gosselin, Hudon-ven der Buhs, Larocque, & Milliard, 2010; Kolb, Wilson, & Taylor, 1992), or (2) underpowered sub-sample sizes for this age group combined with the use of old-fashioned looking, black-and-white facial stimuli (Lawrence et al., 2015). For example, FER developed in a positive linear trend in a forced-choice paradigm in 6- to 16-year-olds, but children between 8 and 13 years old learned no new emotions. Specifically, this age group could identify happiness, sadness, anger and, in contrast to seven-year-olds, surprise with recognition rates above 70%, but not yet disgust and fear in adult faces, which remained around 50% accuracy in a multiple forced-choice task (Lawrence et al., 2015). In a free-labelling task, children between 7 and 11 years old did not reach a 50% accuracy for contempt, whereas roughly two-thirds of those children recognised disgust above this level (Widen & Russell, 2010). Even after early childhood, FER thus seems to develop gradually.

Regarding gender effects, a meta-analysis in children showed a small, but robust female advantage in FER (McClure, 2000). This can possibly be explained by different socialisation of girls and boys, as girls receive more encouragement to develop their emotional intelligence than boys (McClure, 2000; Meadows, 1996). However, later FER-research in children aged 7–13 found no gender differences in recognising anger and fear (Thomas, De Bellis, Graham, & LaBar, 2007). These ambiguous and even contradictory findings may be explained by the use of stimuli with compromised validity for children. By gathering normative data of valid facial stimuli, the current study can help to determine how normal FER develops more precisely. Taken together, it seems important to investigate developmental factors in FER for specific emotional expressions. We therefore decided to investigate the effect of children's gender and age on FER in the current study.

Present study

The current study had two major aims. The first was to validate the adult Caucasian faces of the RaFD for school-aged children. The second aim was to study developmental factors in FER. For readability purposes, the results and discussion section were split in accordance with these two research aims.

Firstly, as a primary validation measure, we investigated the degree of agreement between the intended emotional expression and the emotion children identified in a multiple forced choice task. Additionally, and consistent with Langner et al. (2010), the clarity and valence of the expressed emotions, and the model attractiveness were assessed. Clarity indicates how well an emotion is expressed, valence is one of the most important appraisals of emotional experience (e.g. Scherer, 2013), and attractiveness influences how a face is processed (e.g. Hahn & Perrett, 2014). Therefore, these are important stimulus features for researchers to base their stimulus selection on. Our main hypothesis was that children would accurately recognise emotional expressions, that is, there would be an agreement above chance level between the originally intended emotional expressions and children's choices. This agreement will henceforth be called accuracy because it indicates how accurately children identify displayed expressions. Specifically, we expected that the accuracy would be significantly higher for happiness than for all other emotions (Durand et al., 2007; Gao & Maurer, 2009; Mancini et al., 2013), significantly lower for disgust and fear (Lawrence et al., 2015), and significantly lowest for contempt (Bullock & Russell, 1984; Langner et al., 2010; Widen, Christy, Hewett, & Russell, 2011). Our secondary hypotheses concerned other validation measures. We predicted that children, like adults (see Langner et al., 2010), would rate happiness as most clear and contempt as least clear, in accordance with relative recognition accuracy of those expressions. Furthermore, Lawrence et al. (2015) demonstrated that children between 8 and 13 years old were least accurate in identification of disgust and fear in a set without contempt. Assuming that emotion clarity affects recognition accuracy, we expected children to rate happiness as significantly clearer than all other emotions, disgust and fear as less clear, and contempt as significantly least clear, while clarity ratings of the other expressions were not expected to significantly deviate from the mean. Furthermore, we expected child valence ratings to

reflect a similar pattern as shown in the RaFD validation by adults (Langner et al., 2010), since young children seem to be able to differentiate between emotions of different valence (Bullock & Russell, 1984). Thus, valence was hypothesised to be rated as negative for anger, fear, sadness, disgust and contempt, neutral for neutral and surprise, and positive for happiness. Children were expected to give similar ratings of clarity and valence of the pictures, and of attractiveness of the models, meaning there would be little variation between the ratings of different children on each item, reflected in high intra-class correlation coefficients (ICCs; as was the case for adults in Langner et al., 2010).

Secondly, regarding the developmental factors, we predicted that the accuracy of FER would increase with age (Mancini et al., 2013), and that girls would outperform boys on FER, consistent with the meta-analysis by McClure (2000). By investigating these questions, the current study provides researchers interested in children's responses to emotion expressions with validated stimuli. Thereby, this is the first adult faces database validation study in children. Additionally, this study can advance our understanding of the development of emotion recognition by distinguishing differences in developmental paths for separate emotions, ages and gender.

Methods

The hypotheses, procedure, and analyses were preregistered in the Open Science Framework: https://osf.io/gfk5v/. Stimulus materials are available at www.rafd.nl.

Recruitment and participants

In total, we recruited 15 elementary schools throughout the Netherlands. The current study was conducted in 40 classes of 11 of the 15 schools, as this study was part of a larger project investigating emotion recognition and childhood anxiety. We aimed at children in grade 3–6 (ages 8–13). Parents were asked to sign active informed consent on paper or via an online form. Children indicated if they wanted to participate at the beginning of the sessions, and children of 12 years and older signed active consent as well. Ethical approval was received from the Ethics Committee of the Faculty of Social Sciences (ECSW) from Radboud University, Nijmegen, The Netherlands (reference ECSW2016-2811-447R3001).

The sample size was based on an a priori power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), such that there was an 95% chance to detect a small-to-medium sized effect ($f = .20$) at an alpha level of .05. In total, 554 subjects were needed to investigate interactions between emotion, age and gender effects with sufficient power. After the recruitment and consent procedure, a total of 658 children between 8 and 13 years old participated in this study. This was more than expected, because the number of classes within a school that participated varied highly between schools. Only three children of 13 years old participated. To not generalise findings of such a small sample to a whole age group, they were not included in the analyses. Furthermore, 3 children were excluded from the analyses because they did not participate (due to a dentist appointment, remedial teaching, or later entrance). The final sample included 652 children (330 girls, 319 boys, and 3 children that did not report their gender, $M_{age} = 10.00$, $SD_{age} = 1.26$). Of the final sample, 98% of the children were born in the Netherlands.

Validation task

We used the RaFD straight gaze, front camera pictures of Caucasian adult models (19 female and 20 male models) displaying eight expressions (angry, disgusted, fearful, happy, sad, surprised, contemptuous, and neutral; see Figure 1). Of all 312 pictures, 311 were used due to an accidental omission. To validate all pictures without exhausting the children, 10 slideshow versions, each including 42 pictures, were created in PowerPoint. It started with a block of 10 pictures of 5 male and 5 female models to familiarise the children with the subset and have them rate the attractiveness of these models. The second block consisted of 32 pictures: Two male and two female models from the first block showed all 8 expressions in a randomised order. For each picture, children answered 3 multiple-choice questions in paper answer booklets, respectively: (1) "Which emotion does this person express?" (fear, happiness, sadness, contempt, anger, neutral, surprise, disgust, other); (2) "How clear do you find this emotion?" (5-point Likert scale [1: *not at all* - 5: *very clear*]); (3) "How positive do you find this emotion?" (5-point Likert scale [1: *not at all* - 5: *very positive*]). There were two versions of the answer booklet in which the order of the choice options for the expressed emotion was varied. Version 1 showed the emotion options in the order listed above,

Figure 1. Example expressions from the Radboud Faces Database, from top left: neutral, angry, contemptuous, fearful, happy, disgusted, sad, surprised.

version 2 showed the emotions in reversed order. However, the option *other* was always listed last, since this served to increase response freedom within the forced-choice paradigm. The different versions of the slideshows and booklets were randomly and equally divided across and within the schools, but were kept constant within classrooms. In total, each picture was rated by 49–100 children ($M = 67$, $SD = 16$); each picture was rated by 23–51 girls ($M = 34$, $SD = 9$) and by 22–50 boys ($M = 33$, $SD = 9$).

General procedure

Data gathering took place between March 2017 and June 2017. The study was conducted in classrooms at the participating schools during school hours. The seating arrangement enabled children to clearly see the RaFD pictures on a screen in front of the classroom, with enough space between desks to prevent children from copying each other's answers. All tasks were completed with paper and pencil under supervision of the researchers. At the beginning of the session, children were informed about the tasks, and were reminded that they had the right to stop at all times. The researchers named and explained the specific questions and answering scales that were used, and repeated the explanations if necessary. In the same way, the researchers explained all used emotions in plain language so that children were provided with definitions for all emotions. If children were unsure about the meaning of an emotion, it was explained again. Children were encouraged to ask questions if they had any and were told that they should answer according to their own opinion. The RaFD validation task was completed group-wise, with each picture shown until the last child finished the accompanying questions without time pressure. Next, children individually filled out demographic questions. Overall, the session took approximately one hour per class. Children received a personalised certificate and a thankyou gift. Teachers were offered a school-wise report and a workshop about childhood anxiety.

Preparatory analyses

A univariate analysis of variance (ANOVA) on emotion recognition scores including age in years and gender as factors, showed no differences between children who were vs. were not born in The Netherlands,

between the two versions of the answer booklet, between the different slideshow versions, or between different grades, and no effects of the number of siblings (all $ps > .13$). There was a significant, but small effect of school ($p = .035$, $\eta_p^2 = .03$). Bonferroni-corrected post-hoc tests showed that only one school had marginally significantly lower emotion recognition rates than two other schools ($p = .059$, $p = .097$). No other comparisons were significant. We conducted a generalised linear mixed effects analysis to check if the random variation of the hierarchical data structure of children nested in grades, and grades nested in schools explained a significant part of emotion recognition. This analysis showed that the random factors of school and grades within schools did not have significant effects on emotion recognition ($p = .266$, $p = .187$, respectively). We therefore decided to analyse the sample as a whole and at participant level.

Calculation of accuracy rates of recognizing emotional expressions. First, we calculated raw hit rates: mean percentages of correct responses separately for each emotion category. Raw hit rates signal the accuracy per emotion. We then calculated unbiased hit rates as formulated by Wagner (1993; see also Langner et al., 2010) to take potential biases into account. This procedure corrects for possible answer habits, for example, when participants habitually replace an emotion by one specific other emotion. To calculate unbiased hit rates, we first created a choice matrix with intended and chosen emotional expressions as rows and columns, respectively. Next, the number of ratings in each cell was squared and divided by the product of the marginal values of the corresponding row and column, yielding the unbiased hit rates (Wagner, 1993). The unbiased hit percentages were 5% to 16% below the raw hit percentages per emotion category. In order to correct skewed variances of decimal fractions obtained from counts (Fernandez, 1992), the unbiased hit rates were arcsin-transformed (as recommended by Winer, 1971).

Validation metrics

Results

Recognition accuracy of emotional expressions

Means and standard deviations of all accuracy measures of the total sample are reported in Appendix A, separately for each expression. The overall mean raw hit rate for all children across all emotional expressions was 71.84% ($SD = 12.01$). A repeated-measures ANOVA was performed on the arcsin-transformed unbiased hit rates, with emotion (anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral) as within-subject factor. The assumption of sphericity was violated with epsilon >.75, so the values of within-subject effects were analysed using a Huynh-Feldt correction (Field, 1998; Girden, 1992). This analysis revealed a significant within-subject effect of emotion, $F(6.39, 4157.53) = 341.84$, $p < .001$, $\eta_p^2 = .34$, indicating that some emotions were recognised significantly better than others. Most deviation contrasts from the mean accuracy rate were significant due to the high number of ratings per emotion, so only contrasts with an effect size of $\eta_p^2 \geq .10$ are reported here: As expected, happiness was recognised significantly better than the mean accuracy rate of all emotions ($\eta_p^2 = .77$), while disgust ($\eta_p^2 = .11$), fear ($\eta_p^2 = .25$) and contempt ($\eta_p^2 = .53$) had significantly lower accuracy rates than average. Unbiased hit rates per emotion and significance levels of the deviations from the mean accuracy are displayed in Figure 2.[1]

Pairwise comparisons (Bonferroni corrected) between the emotions showed that the emotion recognition accuracy was higher for happiness than for all other emotions separately ($p < .001$). This was followed by sadness, for which accuracy rates were higher than for all other remaining emotions (for surprise: $p = .055$, for all other emotions: $p < .001$). Next, surprise and neutral both received significantly higher accuracy rates than anger, disgust, fear and contempt (anger versus neutral: $p = .002$, other comparisons: $p < .001$). Anger, in turn, had a (marginally) significantly higher accuracy rate than disgust ($p = .058$), fear and contempt (both $p < .001$). Disgust still had a significantly higher accuracy rate than fear ($p = .005$) and contempt ($p < .001$). Finally, contempt had the lowest accuracy rate (for all comparisons $p < .001$).

Interestingly, for some expressions, incorrect responses were not equally distributed across all other expressions. Three expressions were structurally labelled as three other expressions in significantly more than 11% of the choices, which is the chance level in this choice task with nine options. Specifically, one-sample *t*-tests on all choice rates above 11% showed that children mistakenly identified intended expressions of fear as surprise, and confused intended expressions of contempt with a neutral or *other* expression significantly more often than predicted by chance. Raw hit rates and confusion percentages

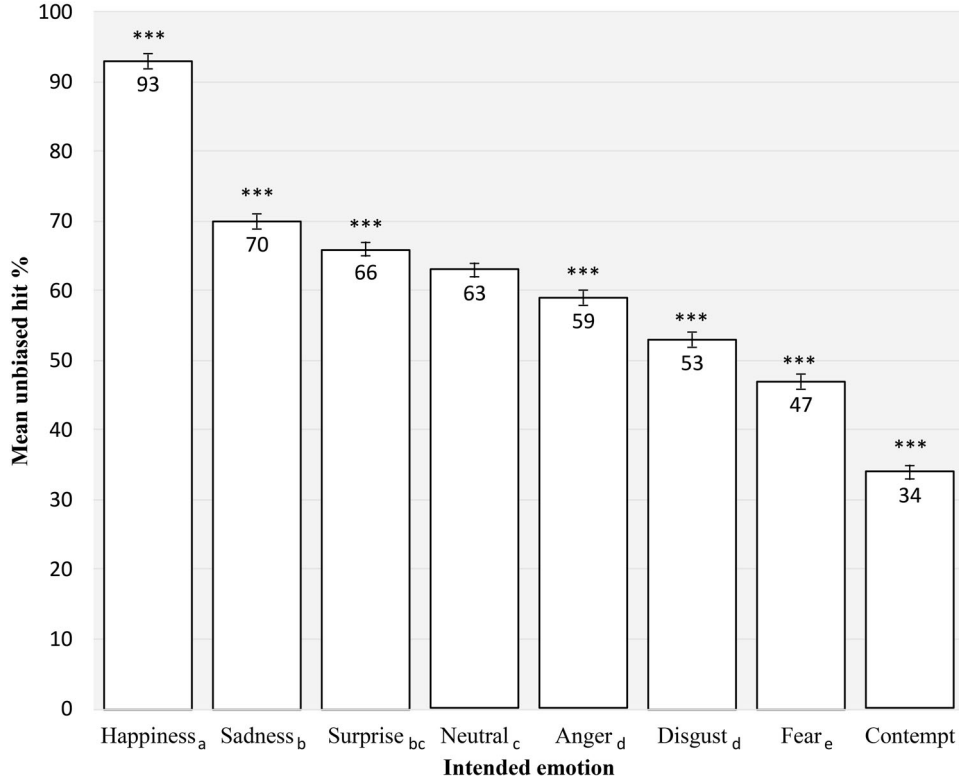

Figure 2. Unbiased accuracy rates per emotion. Error bars represent standard errors. Asterisks indicate significance of deviation from the mean accuracy rate of 60.76 (*SD* = 14.46). Different subscript letters indicate significant differences in accuracy rate between emotions at $p < .01$. $N = 652$. *** $p < .001$.

of all emotions, and accompanying significance levels of differences from chance level are listed in Table 1.

In sum, all emotions were accurately recognised in approximately 72% of the cases. The average FER accuracy for happiness was highest, followed by sadness, surprise and neutral, next by anger, disgust, and fear; and contempt was the hardest to identify, as expected. All differences between accuracy rates were significant, except for the difference between neutral and surprised expressions. Besides, fear was significantly more often labelled as surprise, and contempt as neutral or other than predicted by chance.

Additional validation measures

Interrater reliability indices. Agreement between children on clarity, valence, and attractiveness signals how uniform children's answers were. This was analysed by calculating ICCs for a situation where multiple raters who are presumably representative for other raters, assess each item (Trevethan, 2017). In order to objectively select appropriate ICCs, we calculated the model fit (deviance information criterion) of all possible ICC models. Overall, ICC(2,x)[2] fitted the data best. ICC(2,x) measures for valence, clarity and model attractiveness are reported in Table 2. The current ICCs were all well-above the cut-off for a good degree of similarity and even exceeded the cut-off of ICC(2, k) $\geq .90$ for excellent similarity (Portney & Watkins, 2009), thereby indicating the expected high agreement between children on the validation measures.

Clarity and valence of the pictures. Separate ANOVAs were computed on the average clarity and valence rating per picture with emotion (anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral) as between-subjects factor. Average clarity and valence ratings, and the significance and effect sizes of the deviation from the mean of these ratings are reported in Table 3.

Regarding the clarity ratings, we found that some emotions were perceived as significantly clearer than

Table 1. Emotion choice rates per intended emotion category.

	Chosen emotion								
Intended emotion	Happiness	Surprise	Neutral	Sadness	Anger	Disgust	Fear	Contempt	Other
Happiness	97.89***	0.04	0.49	0.07	0.00	0.10	0.54	0.23	0.66
Surprise	0.44	85.14***	0.96	0.09	0.04	1.04	5.32	2.72	4.27
Neutral	1.65	0.86	79.19***	1.70	2.81	0.69	0.21	4.64	8.26
Sadness	0.19	1.46	3.58	76.89***	0.79	2.86	3.74	5.47	5.00
Anger	0.09	4.66	1.55	3.83	67.74***	5.44	1.14	10.39	5.15
Disgust	0.08	5.18	0.34	0.44	7.52	66.16***	0.39	12.43	7.45
Fear	0.10	20.52**	0.42	1.59	0.91	12.70	51.50***	5.64	6.63
Contempt	4.89	2.35	21.61***	1.45	2.49	4.07	0.31	46.91***	15.92***

Note: The cells contain raw hit percentages. Cells noted with asterisks contain choice percentages above chance level. Values on the main diagonal indicate hits, all others indicate confusions. The hits are marked in green (light grey), significant confusions are marked in red (dark grey).
**$p < .01$. ***$p < .001$.

others, $F(7, 303) = 50.69$, $p < .001$, $\eta_p^2 = .54$. Pairwise comparisons with Bonferroni correction showed that happiness was rated as significantly clearer than all other emotions (all $ps < .001$). Surprise received significantly higher clarity ratings than disgust ($p = .020$), anger ($p = .011$), fear ($p = .003$), neutral, and contempt (both $p < .001$). Finally, contempt was rated as (marginally) significantly less clear than all other emotions (compared to neutral: $p = .068$; compared to the rest: $p < .001$). Furthermore, an exploratory two-sided Pearson correlation test showed that the average clarity ratings per picture correlated significantly and positively with the unbiased hit rates per picture, $r = .73$, $p < .001$, $N = 311$. This indicated that children's ratings of clarity were related to their actual accuracy of emotion recognition.

In sum, the clarity ratings of happiness and surprise were significantly higher than the average of all other emotions, while anger, neutral, fear and contempt had significantly lower clarity scores than the average of all other emotions. As expected, happiness was significantly clearer and contempt significantly less clear compared to all separate emotions.

With regard to emotional valence, an ANOVA revealed that the emotions varied in their valence ratings, $F(7, 303) = 571.25$, $p < .001$, $\eta_p^2 = .93$. Pairwise-comparisons with Bonferroni correction showed that happiness was rated as significantly more positive than all other emotions (all $ps < .001$). Neutral and surprise were rated as significantly more positive than all remaining emotions ($p < .001$ for all comparisons). This was followed by contempt, which was perceived to be significantly less negative than fear, disgust, sadness, and anger ($p < .001$). Of these latter four emotions, fear was significantly less negative than sadness and anger (both $ps < .001$), and disgust was significantly less negative than anger ($p = .007$).

In sum, happiness was seen as most positive, neutral and surprise scored in the middle range of valence, followed by contempt, while fear, disgust, sadness and anger were rated most negatively, as was expected.

Attractiveness of the models. The average attractiveness rating of the models that children gave was 2.04 ($SD = 0.49$). Children assigned both male ($M = 2.00$, $SD = 0.56$) and female ($M = 2.09$, $SD = 0.43$) models with attractiveness rates below the scale mean.

Table 2. Intraclass correlations for picture clarity and valence and model attractiveness.

Dimension	*ICC*(2, 1)[a]	*ICC*(2, *k* = 652)[b]	*ICC*(2, *k* adjusted for missings)[c]
Clarity	0.20	0.99	0.94
Valence	0.42	1.00	0.98
Attractiveness	0.23	0.99	0.98

Note: ICC = intra-class correlation coefficient.
[a]ICC in which one single rating for each item was selected in the calculation;
[b]Indication that the average rating of all raters was taken;
[c]Indication that the average rating of all raters was taken and accounted for missing values.

Table 3. Clarity and valence ratings per emotion.

Intended emotion	Mean clarity (*SD*)	Mean valence (*SD*)
Happiness	4.53 (0.2)***	4.46 (0.2)***
Sadness	3.46 (0.5)	1.82 (0.2)***
Surprise	3.70 (0.5)***	2.66 (0.2)***
Neutral	3.18 (0.3)***	2.71 (0.3)***
Anger	3.36 (0.6)	1.73 (0.2)***
Disgust	3.38 (0.4)	1.92 (0.2)***
Fear	3.33 (0.3)*	2.06 (0.2)***
Contempt	2.89 (0.4)***	2.39 (0.3)*
Average	3.48 (0.6)	2.47 (0.9)

Note: $N = 311$ pictures. Asterisks indicate significance of deviation contrasts from the mean of all ratings of that dimension:
*$p < .05$. **$p < .01$. ***$p < .001$.

Discussion

Within our validation study, we observed that on average, children correctly identified 72% of all emotions. This was in line with our principal hypothesis that children would recognise emotions relatively accurately. Children were able to recognise all emotions above chance level, but hit rates did not equal those of adults (Langner et al., 2010); a two-sample *t*-test of the average hit rates of adults and children showed that they differ significantly ($p < .001$, 95% CI [7.58, 12.42]). For a direct comparison between the recognition rates of the child and adult validation of the RaFD by Langner et al. (2010) and the child validation of the CAFE set by LoBue et al. (2018), see Table 4.

The rank order of accuracy rates per emotion was mostly congruent with the findings of the RaFD validation by adults (Langner et al., 2010), who were most accurate in recognising happiness as well, followed by surprise, anger, neutral, fear, disgust, sadness and contempt, respectively. For adults, only happiness and contempt rates differed significantly from the other expressions, while children also were less accurate in recognising fear and disgust (as expected based on Lawrence et al., 2015). The order of emotions is similar for children and adults, although children especially recognised sadness better in comparison to other emotions than adults did. A possible explanation for this difference is that adults tend to conceal feelings of sadness (Jordan et al., 2011), while children frequently experience and express them (Shipman, Zeman, Nesin, & Fitzgerald, 2003), thus facilitating recognition by accessibility. Furthermore, a study in adults by Recio, Schacht, and Sommer (2013) showed that recognition of sadness was facilitated by static as compared to dynamic portrayal of the emotion. Therefore, the advantage posed by non-moving facial expressions may underlie the relatively high ranking of sadness. The difficulty in recognising fear, disgust and contempt is often found (see, e.g. Rozin, Taylor, Ross, Bennett, & Hejmadi, 2005), and therefore likely to indicate an overall difficulty of interpretation instead of a characteristic of this specific database. The effect of emotion on accuracy rates was large but the adult validation study yielded an even larger effect size (adults: $\eta_p^2 = .52$, current study: $\eta_p^2 = .34$). Even though the adult study showed larger differences between recognition rates of specific emotions, overall the variation in recognition accuracy between emotions is considerable.

Table 4. Direct comparison between the RaFD adult validation 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	98(8)	98(3)	78(42)
Sadness	76(25)	85(16)	49(50)
Surprise	86(22)	90(9)	61(49)
Neutral	79(28)	84(13)	44(50)
Anger	68(25)	81(19)	65(48)
Disgust	66(32)	79(10)	47(50)
Fear	53(32)	88(7)	26(44)
Contempt	49(37)	48(12)	N.A.

Note. RaFD = Radboud Faces Database, CAFE = the Child Affective Facial Expression Set. N.A. = not available. Recognition rates are mean raw accuracy scores and standard deviations per emotion, the scores included from the CAFE set child validation are the scores from the full set at time 1.

Children's emotion misperceptions were similar to the misperceptions of the adult sample tested by Langner et al. (2010) and shown in other FER experiments with child samples (e.g. Gagnon et al., 2010). These confusions can be explained in terms of picture features or conceptual characteristics. First, the finding that fear was mistaken for surprise was found in adults as well, although less often (7%; Langner et al., 2010). The common confusion of fear with surprise in both children (see also: Gagnon et al., 2010; LoBue et al., 2018) and adults (e.g. Goeleven et al., 2008; Langner et al., 2010) might be due to the visual similarities in those two emotions since both involve equal eye-muscle action units: raised eyebrows and upper lids. This explanation has been investigated as the perceptual-attentional limitation hypothesis, proposing that the confusion of fear and surprise comes forth of a failure to perceive differences between the expressions, and if the cues are perceived, not enough attention is directed towards these facial features (Roy-Charland, Perron, Beaudry, & Eady, 2014). A study that manipulated distinctiveness and attention found support for both components of the hypothesis (Chamberland, Roy-Charland, Perron, & Dickinson, 2017), and the current results fit with this explanation as well. Another explanation involves the fact that fear and surprise are conceptually similar. Both emotions are high-arousal states, thereby complicating differentiation (Bullock & Russell, 1984; Jack, Garrod, & Schyns, 2014).

Second, contempt appeared to be difficult to categorise correctly for children, for which multiple possible explanations exist. Conceptually, children often reported that they did not know the meaning of contempt, and they indicated contempt as the most

neutral of all negatively rated emotions, as did adults (Langner et al., 2010). Besides, the only targeted action unit in contemptuous expressions is the uni-lateral lip curl; again, the eye muscles do not change. Taken together, this may have rendered both the concept and portrayal of contempt too ambiguous to recognise, leading to the significantly high number of confusions with a neutral face or choice for the option other. In fact, children were instructed to choose *other* if an expression was not among the choice alternatives, making this a logical choice for an emotion that was either unfamiliar or not clearly expressed. It is not just the case that children are still learning to recognise contempt; adults showed the same confusion with the options *neutral* (Langner et al., 2010; Wagner, 2000) and *other* (Langner et al., 2010). The current findings thus emphasise that contempt expressed by merely a uni-lateral lip curl is easily confused with other expressions, either because of the subtle display (see also Wagner, 2000) or because the label contempt is difficult to use correctly (Matsumoto & Ekman, 2004). Both the confusion of fear and surprise and the ambiguity of contempt have been shown in multiple studies with static and dynamic stimuli (see, e.g. Jack et al., 2014; Wingenbach, Ashwin, & Brosnan, 2016), thereby suggesting that the RaFD has been successfully validated by children.

Concerning the secondary validation measures, clarity ratings roughly mirrored the FER accuracy rates of different emotions. As hypothesised, children found expressions of happiness significantly most clear, and contemptuous faces significantly less clear than average. This pattern indicates that children were able to estimate their ability to recognise emotions, a finding that is supported by the positive correlation between accuracy and clarity. In line with Langner et al. (2010) and our expectations about valence, happiness was perceived as most positive, neutral and surprise as neutral, and the other emotions as negative. This confirms that the emotional valence was perceived as intended, demonstrating the validity of the database. Finally, as hypothesised, children highly agreed with each other on attractiveness, valence, and clarity ratings. This similarity of ratings indicates that the attributes of the items can reliably be used for other raters as well. Consequently, researchers may use valence, clarity or attractiveness as selection criteria or manipulated features in their studies. However, some clarity ratings were different from our hypotheses. Unexpectedly, surprise was rated as significantly clearer than other emotions. An explanation for this finding might be that surprise was the only expression involving a dropped jaw (Langner et al., 2010). This changes the overall shape of the face, making it easier to distinguish it from other emotions (Zou, Ji, & Nagy, 2007). Accordingly, Lawrence et al. (2015) also found that children in middle childhood could correctly identify surprise, although in our sample, children overgeneralised by also confusing fear with surprise. Another finding contrasting our expectations was that disgust and fear were not rated as significantly less clear than average, while they were less accurately recognised. However, the respective means were in the expected direction, below the average clarity rating.

Overall, the current results show that children have no problems interpreting emotional valence, but do have typical difficulties in distinguishing emotions with similar alternatives in the choice-paradigm, bearing implications for the applicability of the RaFD adult pictures in child research. The gap with adult FER accuracy can be explained by the still ongoing development of visual abilities in puberty (Gao & Maurer, 2009), and by the fact that compared to adults, children have had less perceptual exposure to emotional expressions (Pollak, Messner, Kistler, & Cohn, 2009). Also, children have less experience in enacting emotions, while this experience in the form of mimicry has been shown to facilitate explicit emotion recognition (Conson et al., 2013). The hit rates showing differentiation between positive and negative emotions, but confusions between conceptually similar emotions, are in line with Bullock and Russell's (1984) differentiation account and the perceptual-attentional limitation hypothesis (Roy-Charland et al., 2014). This implies that when creating stimulus materials for school-aged children, pictures of emotions that are highly similar either conceptually or in how they are physically expressed (e.g. fear and surprise, contempt and neutral), should best not be combined within one study if the goal is to investigate responses to each of these emotions in specific, because children may confuse the expressions.

Developmental variables in relation to FER

Results

Means and standard deviations of each emotion, separately for gender and age, can be found in Table 5.

Table 5. Mean unbiased hit percentages per participant age group and gender for each emotion category.

		Intended emotion							
Age	Participant gender	Happiness *M* (*SD*)	Sadness *M* (*SD*)	Surprise *M* (*SD*)	Neutral *M* (*SD*)	Anger *M* (*SD*)	Disgust *M* (*SD*)	Fear *M* (*SD*)	Contempt *M* (*SD*)
8	Girls ($n = 53$)	95(10)	78 (22)	60 (25)	50 (34)	63 (27)	52 (33)	37 (33)	22 (25)
	Boys ($n = 35$)	92 (15)	61 (25)	60 (27)	50 (31)	48 (29)	38 (30)	45 (33)	25 (30)
9	Girls ($n = 83$)	92 (15)	71 (26)	64 (25)	60 (30)	61 (26)	50 (29)	44 (33)	34 (32)
	Boys ($n = 84$)	93 (14)	68 (29)	53 (30)	53 (30)	51 (27)	49 (33)	44 (30)	27 (20)
10	Girls ($n = 63$)	95 (11)	73 (28)	67 (25)	66 (30)	61 (34)	58(31)	44 (29)	40 (30)
	Boys ($n = 79$)	94 (15)	65 (27)	63 (27)	66 (27)	59 (24)	50 (31)	49 (34)	41 (36)
11	Girls ($n = 84$)	95 (10)	73 (24)	73 (26)	74 (25)	63 (25)	55 (29)	51 (29)	37 (32)
	Boys ($n = 81$)	92 (14)	70 (25)	72 (26)	66 (26)	60 (28)	60 (34)	54 (30)	36 (34)
12	Girls ($n = 47$)	92 (13)	74 (24)	76 (21)	71 (27)	65 (25)	57 (29)	50 (30)	39 (30)
	Boys ($n = 40$)	94 (12)	69 (31)	68 (24)	72 (20)	56 (22)	52 (31)	47 (28)	32 (28)

Overall means per age group for boys and girls separately are shown in Figure 3. We performed a repeated-measures ANOVA with emotion (anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral) as within-subjects factor, and gender (girl, boy) and age (8, 9, 10, 11, 12) as between-subjects factors on the arcsin-transformed unbiased hit rates. The three children that did not report their gender were excluded from this analysis. Since sphericity was violated with epsilon >.75, within-subject effects were adjusted using a Huynh-Feldt correction (Field, 1998; Girden, 1992).[3] As expected, this test yielded a main

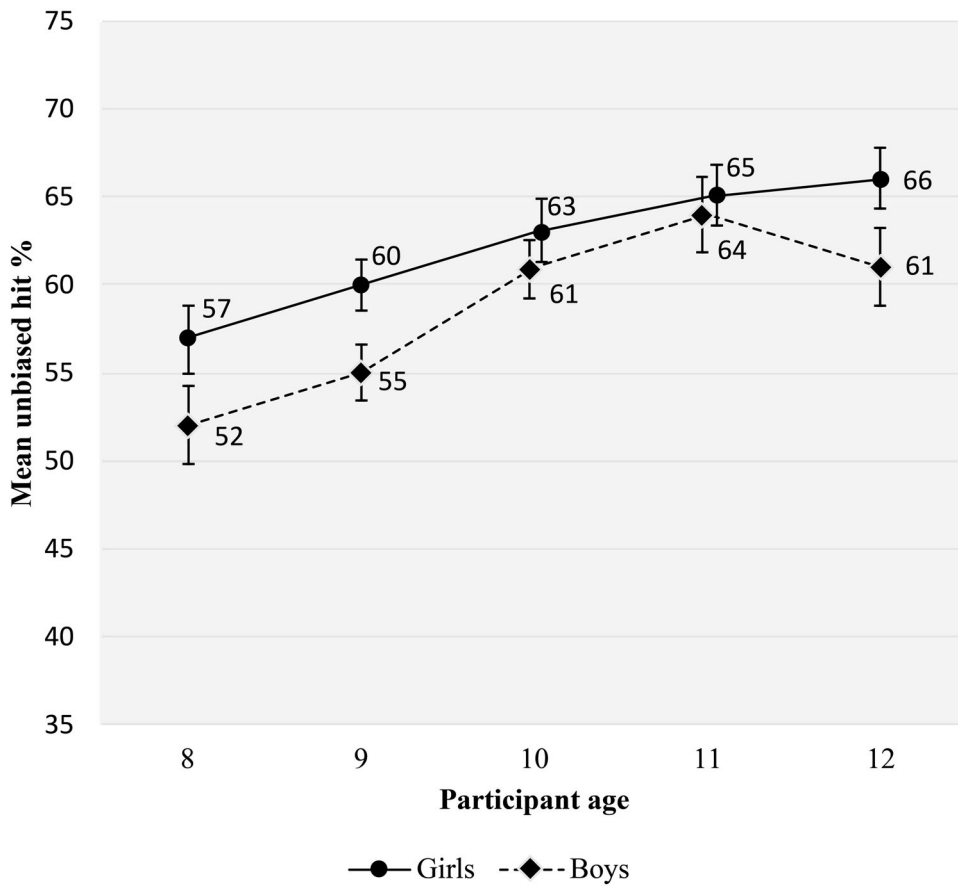

Figure 3. Average unbiased accuracy rate per age group and gender. Error bars represent standard errors. Note that age was assessed in whole years, so the lines connecting data points illustrate the average change, not the measurement of a slope.

effect of emotion on the accuracy rate, $F(6.48, 4140.15) = 316.00$, $p < .001$, $\eta_p^2 = .33$, which was similar to the pattern yielded by the repeated-measures ANOVA without between-subjects factors (see page 6 of this paper).

Age effects. Children of different ages varied significantly in how well they recognised emotions, $F(4, 639) = 8.77$, $p < .001$, $\eta_p^2 = .05$. Pairwise comparisons of ages with Bonferroni-correction showed that: 8-year-old children ($M_{\text{raw hit rate}} = 67\%$, $SD = 12\%$) had significantly lower accuracy rates compared to 10- ($M_{\text{raw hit rate}} = 73\%$, $SD = 12\%$; $p = .009$), 11- ($M_{\text{raw hit rate}} = 75\%$, $SD = 11\%$; $p < .001$) and 12- ($M_{\text{raw hit rate}} = 75\%$, $SD = 11\%$; $p = .005$) year-old children. Next, 9-year-olds ($M_{\text{raw hit rate}} = 69\%$, $SD = 13\%$) had lower accuracy rates than 11- and 12-year-olds ($p < .001$ and $p = .055$, respectively). The children of 10, 11, and 12 years old did not significantly differ in accuracy rates. The interaction between age and emotion on arcsin-transformed accuracy rates was significant, $F(25.92, 4140.15) = 1.59$, $p = .03$, $\eta_p^2 = .01$. This means that the increase in FER accuracy with age varied slightly between emotions. Specifically, parameter estimates showed that 8-year-olds had significantly lower accuracy on recognising contempt ($B = -.21$, $p = .016$, $SE = .09$), and 8- and 9-year-olds had a lower accuracy for recognising neutral faces (respectively $B = -.29$, $p = .003$, $SE = .10$; $B = -.18$, $p = .039$, $SE = .09$) and surprised faces (8-year-olds: $B = -.27$, $p = .003$, $SE = .09$; 9-year-olds: $B = -.18$, $p = .031$, $SE = .08$), as compared to 12-year-olds. The interaction between age and gender was non-significant ($p = .86$), suggesting that the accuracy rates for children of different ages did not vary significantly between boys and girls.

Gender effects. Overall, girls ($M_{\text{raw hit rate}} = 73\%$, $SD = 11\%$) were slightly more accurate than boys ($M_{\text{raw hit rate}} = 71\%$, $SD = 13\%$), $F(1, 639) = 8.50$, $p = .004$, $\eta_p^2 = .01$. There was also a small, but significant interaction between emotion and gender on accuracy rates, $F(6.48, 4140.15) = 3.01$, $p = .005$, $\eta_p^2 = .01$, indicating that some emotions were significantly better recognised by boys and others by girls. Although average accuracy rates for each emotion separately were higher among girls than boys (except for fear), this higher accuracy of girls only reached marginal significance for anger ($B = -.17$, $p = .057$, $SE = .09$). All other comparisons did not statistically differ (all $ps > .10$). Lastly, the three-way interaction between emotion, gender and age was not significant ($p = .155$), indicating that boys' and girls' patterns of accuracy on the different emotional expressions did not vary with age.

In sum, as hypothesised, FER differed with age, but not between all age groups. Eight-year-olds chose the intended emotion significantly less often than children of 10 and older, and nine-year-olds were only significantly outperformed by 11- and 12-year-old children. The small, significant interaction between age and emotion indicated particular increases in accuracy for surprised, contemptuous and neutral expressions. Overall, as expected, girls had significantly higher agreement rates than boys, but this effect was small, and the interaction between gender and emotion showed that this difference was only marginally significant for anger and not significant for the other emotions.

Discussion

The second study aim was to assess age and gender differences within our sample. As hypothesised, in general, older children had a higher emotion recognition accuracy than younger children. In specific, 10- to 12-year-olds were more accurate than 8- and 9-year-olds. The basic FER skills of 8-year-olds confirm that FER development starts earlier in childhood, which is in accordance with the findings of LoBue et al. (2018), but the variation of FER with age indicates some development of FER in late childhood as well.

The most obvious increase in emotion accuracy took place around the age of 9 and 10. This coincides with neurological findings by Batty and Taylor (2006), who showed that a specific brain activity (event-related potential component N170) in processing facial emotions was modulated and became faster for children of 10 years and older compared to 9 years and younger. Meaux et al. (2014) explained this change in neurological activity by increased early attention for the eye-region in children between 6 and 10 years of age. The eyes are an important area for holistic face processing (Rossion, 2013), and a recent experiment on face identification showed that children are more dependent on holistic face processing than adults (Billino, van Belle, Rossion, & Schwarzer, 2018). FER also is more accurate if faces are processed holistically (Batty & Taylor, 2006). Schwarzer (2000) showed that 10-year-olds are more likely to process faces holistically compared to 7-year-olds, which may indicate that children develop more efficient ways of processing facial stimuli around the age of 10. Our support for a non-linear, stepwise progress in FER performance of school-

aged children emphasises the need for research investigating a wider age range. This may provide clarity about how FER develops through childhood and into adulthood, and how this is linked to face identification in general.

This development with age contrasts the stagnation shown by Lawrence et al. (2015), but confirms findings from studies with smaller but varying age ranges (Lee, Quinn, Pascalis, & Slater, 2013). The discrepancies between studies may result from different emotion selections, since the current results showed the sharpest increase in recognition of neutral, contemptuous and surprised faces. For instance, Lawrence et al. (2015) did not include neutral and contemptuous expressions. The observed interaction between emotion and age should be interpreted with care due to the small effect size. One explanation of the interaction could be that neutrality, contempt and surprise are more complicated facial expressions (Jack et al., 2014), which children only later learn to process and distinguish (Bullock & Russell, 1984). Surprise, neutrality and contempt had the most ambiguous valence in the present study, making them more complex to distinguish than clearly negative or positive emotions, especially for the younger children (Bullock & Russell, 1984). Furthermore, the increase in recognising contempt may go hand in hand with language development (Barrett, Lindquist, & Gendron, 2007). Even though the emotion label was explained to children that were not familiar with it, personal experience with the verbal label may have helped older children to correctly categorise it.

Furthermore, in line with our hypothesis about gender effects, we found a small FER advantage for girls. Given the large sample size, the significance of the small gender effects should be interpreted with care. The exact gender effect varied per emotion; girls were marginally significantly better than boys in recognising anger, but the girls' higher average accuracy was not significant for the other emotions. The current effect size was as small as the effect found by Lawrence and colleagues, although their study showed significantly higher accuracy rates for girls on surprise and disgust. In contrast, Thomas et al. (2007) did not find a difference in FER between boys and girls, and the current results suggest that this could be due to their selection of emotions or to the difference between processing static versus dynamic facial expressions. On the one hand, Thomas and colleagues only investigated recognition of anger and fear, while in the current study, fear was the only emotion with a higher accuracy (though insignificant) for boys. On the other hand, we used static emotional expressions while Thomas et al. used morphs, and girls may have a specific advantage in recognising static emotional images. Finally, their study included only 31 children, which may have been too few to reveal a small effect. Overall, this asks for further, well-powered research into the mechanisms behind gender FER differences, differentiating between emotions.

General discussion

The aim of this study was twofold. Our first main goal was to validate the Caucasian adult pictures of the RaFD from a child's perspective. Specifically, we tested to what extent children between 8 and 12 years old could accurately identify expressed emotions. Additionally, we assessed the clarity and valence of the pictures, as well as the models' attractiveness. Our second main goal was to investigate age and gender effects on FER. The results showed subtle advantages for older children and for girls.

The current study has several strengths, among them the well-powered, pre-registered sample size and the assessment of multiple validation measures per picture. However, several limitations need to be mentioned. First, the current design could not determine to what extent the accuracy rates on expressions were due to the FER of children or the quality of the stimuli. However, the comparable results from studies using other databases (e.g. Lawrence et al., 2015; Wagner, 2000) imply that developmental variables and emotion-specific characteristics determined the pattern of FER, meaning that the RaFD pictures depict the emotions as intended. Second, conducting the study class-wise created practical limitations. For instance, although children were explicitly told to ask questions if anything was unclear, they may have felt uncomfortable asking questions in front of their class mates, and conceptual understanding of some emotions may have been lacking, compromising reliability of the hit rates. Class-wise task administration also meant that children who responded faster had to wait, and boredom might have led to attention lapses (Malkovsky, Merrifield, Goldberg, & Danckert, 2012). Children who worked slower might have felt hurried, even though no time limits were imposed. This may have negatively biased the results, so future research might benefit from an individually

administered validation task. Contrarily, the forced-choice paradigm may have led to over-estimation of children's ability to recognise some emotions (Wagner, 2000), although scores were unbiased and the option *other* was included to minimise these effects. In a methodological study, Frank and Stennett (2001) assessed the effectiveness of adding *other* to the answer options in facial expression choice tasks and showed that this modified format counteracts the bias that is present in mere forced-choice tasks (Russell, 1993). Finally, some questions remain to be answered by future research. For instance, what is the exact role of visual versus cognitive ability in FER performance? Can RaFD stimuli be used in children under 8 as well? Also, the differences between adults and children in recognising emotions on adult faces give rise to the question how accurately children can identify emotions of the RaFD child models, given the potential benefit of own-age effects in child faces. Validating the remaining RaFD stimuli of child faces and Moroccan faces in children would enhance the utility of the database even further.

Future prospects and conclusions

Until now, most studies with child participants did not have access to facial emotional stimuli validated by children. This study provides researchers with child validation data of a large, standardised database. Together with the study by LoBue et al. (2018), this research facilitates the use of validated face databases in (developmental) studies. The somewhat lower FER rates of school-aged children compared to adults emphasise the importance of separate validation data for this age group. With an overall FER rate of 72%, our data lend support to the validity of the RaFD Caucasian adult pictures to use in experiments with 8- to 12-year-old children. Besides the overall high recognition, the stimuli are especially promising as valence signals, and can be used to depict specific emotions as long as researchers are aware of possible confusions between highly similar emotions (please see page 6–7 for the exact confusions). Specific pictures can be selected based on the validation data, which are freely available online (see: https://osf.io/gfk5v/files/). To name some examples, developmental, social, clinical and neurological research may use pictures to manipulate emotional dimensions, to control for possible confounds, or to gather normative data with a sharp eye on matching gender and age groups. Next to these practical uses, theoretical FER research can incorporate a broader variety of emotions in children, because the current findings include outcomes concerning contempt. This is relevant because the differences between emotions, as shown in this study, imply that conclusions about children's FER development depend on the emotions. Future FER research in children thus needs attention to subtle differences in gender, age and emotion, and it can benefit from this study by using the child validation data.

Notes

1. Please note that, to facilitate interpretation, all figures display unbiased hit percentages instead of the transformed arcsine radians used in the statistical analyses.
2. The basis for this measure was a two-way ANOVA (Shrout & Fleiss, 1979), which is typically cross-classified: cells are identifiable by row and column indices. To match this theoretical assumption, cross-classified multilevel Markov Chain Monte Carlo ICC's were calculated. These are not used very often, but were chosen here because of the best data-fit.
3. As an additional check for the robustness of the results, a linear mixed model was conducted with the arcsin-transformed unbiased hit rates as dependent variable, emotion as fixed repeated-measures factor, age and gender as fixed between-subject factors, and including all interactions. A restricted maximum likelihood (REML) covariance matrix estimation was applied to control for potential violations of the model assumptions. All pairwise comparisons were Bonferroni-corrected for multiple testing. This analysis showed the same significant effects as the pre-registered repeated-measures ANOVA, which is described here.

Acknowledgements

We thank the children and their parents and the schools for participating. Also, we thank Giovanni ten Brink for his contribution to the data preparation and Pierre Souren for calculating the interrater reliability indices. Lastly, we thank the reviewers for their helpful comments on a previous version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Iris A. M. Verpaalen http://orcid.org/0000-0002-4168-7545
Lynn Mobach http://orcid.org/0000-0002-0172-8525

References

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, *11*(8), 327–332. doi:10.1016/j.tics.2007.06.003.

Batty, M., & Taylor, M. J. (2006). The development of emotional face processing during childhood. *Developmental Science*, *9*(2), 207–220. doi:10.1111/j.1467-7687.2006.00480.x.

Billino, J., van Belle, G., Rossion, B., & Schwarzer, G. (2018). The nature of individual face recognition in preschool children: Insights from a gaze-contingent paradigm. *Cognitive Development*, *47*, 168–180. doi:10.1016/j.cogdev.2018.06.007.

Brown, H. M., Eley, T. C., Broeren, S., MacLeod, C., Rinck, M. H. J. A., Hadwin, J. A., & Lester, K. J. (2014). Psychometric properties of reaction time based experimental paradigms measuring anxiety-related information-processing biases in children. *Journal of Anxiety Disorders*, *28*(1), 97–107. doi:10.1016/j.janxdis.2013.11.004.

Bullock, M., & Russell, J. A. (1984). Preschool children's interpretation of facial expressions of emotion. *International Journal of Behavioral Development*, *7*(2), 193–214. doi:10.1177/016502548400700207.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641–651. doi:10.1038/nrn172.

Chamberland, J., Roy-Charland, A., Perron, M., & Dickinson, J. (2017). Distinction between fear and surprise: an interpretation-independent test of the perceptual-attentional limitation hypothesis. *Social Neuroscience*, *12*(6), 751–768. doi:10.1080/17470919.2016.1251964

Conson, M., Ponari, M., Monteforte, E., Ricciato, G., Sarà, M., Grossi, D., & Trojano, L. (2013). Explicit recognition of emotional facial expressions is shaped by expertise: Evidence from professional actors. *Frontiers in Psychology*, *4*), doi:10.3389/fpsyg.2013.00382.

Corbett, B. A., Key, A. P., Qualls, L., Fecteau, S., Newsom, C., Coke, C., & Yoder, P. (2016). Improvement in social competence using a randomized trial of a theatre intervention for children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*(2), 658–672. doi:10.1007/s10803-015-2600-9.

De Sonneville, L. M. J., Verschoor, C. A., Njiokiktjien, C., Op het Veld, V., Toorenaar, N., & Vranken, M. (2002). Facial identity and facial emotions: Speed, accuracy, and processing strategies in children and adults. *Journal of Clinical and Experimental Neuropsychology*, *24*(2), 200–213. doi:10.1076/jcen.24.2.200.989.

Durand, K., Gallay, M., Seigneuric, A., Robichon, F., & Baudouin, J. Y. (2007). The development of facial emotion recognition: The role of configural information. *Journal of Experimental Child Psychology*, *97*(1), 14–27. doi:10.1016/j.jecp.2006.12.001.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES - A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, *42*(1), 351–362. doi:10.3758/BRM.42.1.351.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi:10.3758/BF03193146.

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. Retrieved from: http://www.discoveringstatistics.com/docs/sphericity.pdf

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. doi: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. doi:10.1007/s10919-009-0076-z.

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. doi:10.1016/j.jecp.2008.11.002.

Gervais, M. M., & Fessler, D. M. (2017). On the deep structure of social affect: Attitudes, emotions, sentiments, and the case of "contempt". *Behavioral and Brain Sciences*, *40*), doi:10.1017/S0140525X16000352.

Girden, E. R. (1992). *ANOVA: Repeated measures. Sage university paper series on quantitative applications in the social sciences, 07-084*. Newbury Park, CA: Sage.

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska directed emotional faces: A validation study. *Cognition & Emotion*, *22*(6), 1094–1118. doi:10.1080/02699930701626582.

Grossmann, T., & Johnson, M. H. (2007). The development of the social brain in human infancy. *The European Journal of Neuroscience*, *25*(4), 909–919. doi:10.1111/j.1460-9568.2007.05379.x.

Hahn, A. C., & Perrett, D. I. (2014). Neural and behavioral responses to attractiveness in adult and infant faces. *Neuroscience & Biobehavioral Reviews*, *46*, 591–603. doi:10.1016/j.neubiorev.2014.08.015.

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192. doi:10.1016/j.cub.2013.11.064.

Jordan, A. H., Monin, B., Dweck, C. S., Lovett, B. J., John, O. P., & Gross, J. J. (2011). Misery has more company than people think: Underestimating the prevalence of others' negative emotions. *Personality and Social Psychology Bulletin*, *37*(1), 120–135. doi:10.1177/0146167210390822.

Kolb, B., Wilson, B., & Taylor, L. (1992). Developmental changes in the recognition and comprehension of facial expression: Implications for frontal lobe function. *Brain and Cognition*, *20*(1), 74–84. doi:10.1016/0278-2626(92)90062-Q.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud faces database. *Cognition & Emotion*, *24*(8), 1377–1388. doi:10.1080/02699930903485076.

Lawrence, K., Campbell, R., & Skuse, D. (2015). Age, gender, and puberty influence the development of facial emotion

recognition. *Frontiers in Psychology*, *6*(June), 761. doi:10.3389/fpsyg.2015.00761.

Lee, K., Quinn, P. C., Pascalis, O., & Slater, A. (2013). Development of face processing ability in childhood. In P. D. E. Zelaso (Ed.), *The Oxford Handbook of developmental psychology, Vol 1: Body and Mind* (pp. 338–370). New York: Oxford University Press.

Leppänen, J. M. (2011). Neural and developmental bases of the ability to recognize social signals of emotions. *Emotion Review*, *3*(2), 179–188. doi:10.1177/1754073910387942.

LoBue, V., Baker, L., & 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. doi:10.1080/02699931.2017.1365046.

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

Mancini, G., Agnoli, S., Baldaro, B., Bitti, P. E. R., & Surcinelli, P. (2013). Facial expressions of emotions: Recognition accuracy and affective reactions during late childhood. *The Journal of Psychology*, *147*(6), 599–617. doi:10.1080/00223980.2012.727891.

Matsumoto, D., & Ekman, P. (2004). The relationship among expressions, labels, and descriptions of contempt. *Journal of Personality and Social Psychology*, *87*(4), 529–540. doi:0.1037/0022-3514.87.4.529.

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin*, *126*(3), 424–453. doi:10.1037/0033-2909.126.3.424.

Meadows, S. (1996). *Parenting behavior and child's cognitive development*. London, UK: Psychology Press.

Meaux, E., Hernandez, N., Carteau-Martin, I., Martineau, J., Barthélémy, C., Bonnet-Brilhault, F., & Batty, M. (2014). Event-related potential and eye tracking evidence of the developmental dynamics of face processing. *European Journal of Neuroscience*, *39*(8), 1349–1362. doi:10.1111/ejn.12496.

Miceli, M., & Castelfranchi, C. (2018). Contempt and disgust: The emotions of disrespect. *Journal for the Theory of Social Behaviour*, *48*(2), 205–229. doi:10.1111/jtsb.12159.

Pollak, S. D., Messner, M., Kistler, D. J., & Cohn, J. F. (2009). Development of perceptual expertise in emotion recognition. *Cognition*, *110*(2), 242–247. doi:10.1016/j.cognition.2008.10.010.

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice* (3rd ed). Upper Saddle River, NJ: Pearson Education.

Recio, G., Schacht, A., & Sommer, W. (2013). Classification of dynamic facial expressions of emotion presented briefly. *Cognition & Emotion*, *27*(8), 1486–1494. doi:10.1080/02699931.2013.794128.

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

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253. doi:10.1080/13506285.2013.772929.

Roy-Charland, A., Perron, M., Beaudry, O., & Eady, K. (2014). Confusion of fear and surprise: A test of the perceptual-attentional limitation hypothesis with eye movement monitoring. *Cognition & Emotion*, *28*, 1214–1222. doi:10.1080/02699931.2013.878687.

Rozin, P., Taylor, C., Ross, L., Bennett, G., & Hejmadi, A. (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. doi:10.1080/02699930441000166.

Russell, J. A. (1993). Forced-choice response format in the study of facial expression. *Motivation and Emotion*, *17*(1), 41–51. doi:10.1007/BF00995206.

Scherer, K. R. (2013). The nature and dynamics of relevance and valence appraisals: Theoretical advances and recent evidence. *Emotion Review*, *5*(2), 150–162. doi:10.1177/1754073912468166

Schwarzer, G. (2000). Development of face processing: The effect of face inversion. *Child Development*, *71*(2), 391–401. doi:10.1111/1467-8624.00152.

Shipman, K. L., Zeman, J., Nesin, A. E., & Fitzgerald, M. (2003). Children's strategies for displaying anger and sadness: What works with whom? *Merrill-Palmer Quarterly*, *49*(1), 100–122. doi:10.1353/mpq.2003.0006.

Shrout, P. E., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi:10.1037/0033-2909.86.2.420.

Thomas, L. A., De Bellis, M. D., Graham, R., & LaBar, K. S. (2007). Development of emotional facial recognition in late childhood and adolescence. *Developmental Science*, *10*(5), 547–558. doi:10.1111/j.1467-7687.2007.00614.x.

Trevethan, R. (2017). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127–143. doi:10.1007/s10742-016-0156-6.

Van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, *11*(4), 907–920. doi:10.1037/a0023853.

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

Wagner, H. L. (2000). The accessibility of the term "contempt" and the meaning of the unilateral lip curl. *Cognition & Emotion*, *14*(5), 689–710. doi:10.1080/026999300501176675.

Widen, S. C., Christy, A. M., Hewett, K., & Russell, J. A. (2011). Do proposed facial expressions of contempt, shame, embarrassment, and compassion communicate the predicted emotion? *Cognition & Emotion*, *25*(5), 898–906. doi:10.1080/02699931.2010.508270.

Widen, S. C., & Russell, J. A. (2010). Children's scripts for social emotions: Causes and consequences are more central than are facial expressions. *British Journal of Developmental Psychology*, *28*, 565–581. doi:10.1348/026151009X457550d.

Winer, B. (1971). *Statistical principles in experimental design* (2nd ed). New York: McGraw-Hill.

Wingenbach, T. S. H., Ashwin, C., & Brosnan, M. (2016). Validation of the Amsterdam Dynamic Facial Expression Set - Bath Intensity Variations (ADFES-BIV): A Set of Videos Expressing Low, Intermediate, and High Intensity Emotions. *PloS One*, *11*(1), e0147112. doi:10.1371/journal.pone.0147112.

Zou, J., Ji, Q., & Nagy, G. (2007). A comparative study of local matching approach for face recognition. *IEEE Transactions on Image Processing*, *16*(10), 2617–2628. doi:10.1109/TIP.2007.904421.

Appendix

Overview of Mean Hit Rates per Emotion, Analysed per Participant

	Intended emotion							
Type of hit rate	Happiness *M* *(SD)*	Sadness *M* *(SD)*	Surprise *M* *(SD)*	Neutral *M* *(SD)*	Anger *M* *(SD)*	Disgust *M* *(SD)*	Fear *M* *(SD)*	Contempt *M* *(SD)*
Raw %	98(8)	76(25)	86(22)	79(28)	68(25)	66(32)	53(32)	49(37)
Unbiased %	93(13)	70(26)	66(26)	63(29)	59(26)	53(31)	47(31)	34(32)
Arcsin-transformed	1.4(0.3)	0.9(0.5)	0.8(0.5)	0.8(0.5)	0.7(0.4)	0.6(0.5)	0.6(0.5)	0.4(0.4)