Children’s Perception of Facial Expressions

Pieter F. de Bordes1, Fred Hasselman2, and Ralf F. A. Cox1
1 Department of Psychology, University of Groningen
2 Behavioural Science Institute, Radboud University

This study investigated the developing ability of children to identify emotional facial expressions in terms of the contexts in which they generally occur. We presented Dutch 6- to 9-year-old primary school children (N = 164, 98 girls) prototypical contexts for different emotion categories and asked them whether different kinds of facial expressions belonged to those contexts or not, using a 2-alternative forced-choice task. Correct and incorrect responses were quantified into a single index using signal detection theory, representing children’s sensitivity to perceive each facial expression as categorically different from each of the others in terms of their prototypical contexts. Results show age-related improvements in identifying facial expressions as belonging to their prototypical contexts. In addition, we found that older children not only made less misidentifications but also misidentified less kinds of facial expressions to the prototypical contexts. Furthermore, the kinds of misidentifications children made suggest that they do not identify facial expressions based on their conceptual emotional valence. Results were discussed from a perceptual learning account.

Keywords: facial expressions, signal detection theory, Shannon Diversity Index, social development, perceptual learning

Supplemental materials: https://doi.org/10.1037/dev0000979.supp

Humans are affected by facial expressions of others and tend to adjust their behavior accordingly (Gao & Maurer, 2009; Leppänen & Nelson, 2009; Zebrowitz, 2011). The ability to perceive distinct facial expressions is crucial for social adaption and social exchange (Bal et al., 2010; Gao & Maurer, 2010; Mostow et al., 2002). This ability starts to develop from infancy (Cong et al., 2019; Lee et al., 2015) and continues throughout childhood and adolescence (Herba & Philips, 2004; Montirozzo et al., 2010; van Beek & Dubas, 2008; Widen & Russell, 2003). Although this development has been well documented, research outcomes and suggested developmental pathways seem to vary according to how the ability is operationalized and theoretically conceived (Calvo & Marrero, 2009; Durand et al., 2007; Rodger et al., 2015, 2018; Vicari et al., 2000; Widen & Russell, 2008). In the current article, we will argue that the perception of facial expressions is contextualized. In addition, we present a novel method of analyzing the development of children to identify facial expressions as belonging to the prototypical contexts in which they are normally perceived.

Measuring Categorical Perception of Facial Expressions

When it comes to investigating the development of children in the ability to perceive facial expressions, a wide range of research methods have been used such as labeling tasks, discrimination tasks, and recognition tasks (e.g., Herba et al., 2006; Rodger et al., 2018; Thomas et al., 2007; Vicari et al., 2000; Widen, 2013). Most of these methods aim to measure the developing ability of children to perceive each facial expression in isolation (e.g., Durand et al., 2007; Rodger et al., 2015), in reference to only neutral expressions (e.g., Gao & Maurer, 2009, 2010) or in reference to just a few other categories of expressions (e.g., Thomas et al., 2007). This might offer a limited view on their ability to perceive distinct facial expressions for two reasons. First, the number and kinds of categories of facial expressions available during testing likely affect how children categorize the perceived facial expressions (Bimler & Kirkland, 2001; Russell & Fehr, 1987). Second and more important, each discrete category of facial expressions is unlikely to be learned in isolation. This is because being able to perceive a particular facial expression as a discrete category entails that it is perceived as different from other categories of facial expressions. For instance, being able to perceive facial expressions of anger entails the ability to perceive a happy facial expression as...
“not angry.” To overcome these limitations, it seems fruitful to use a forced-choice task in which children need to identify different facial expressions as belonging to an emotion category and to analyze both the correct and the incorrect responses. This can reveal to what extent children perceive facial expressions as categorically discrete and mutually exclusive. Based on such an analysis, Widen and Russell (2003, 2008) found that although 5-year-old children are not yet proficient in verbally labeling expressions of anger accurately, they rarely mislabeled expressions of anger as happy or sad. This shows that although young children might not have the ability to label angry expressions properly, they are able to exclude angry expressions from other discrete categories, which is at least equally important in terms of attaching meaning to the perceived angry expression. In addition, they found that whereas 4-year-old children used the label “happy” to include happy, surprised, and fearful expressions, 5- to 6-year-olds would be less inclined to mislabel surprised and fearful expressions as “happy” and were able to label those expressions appropriately (Widen & Russell, 2003, 2008). In general, Widen and Russell (2003, 2008) have found that the categories of facial expressions emerge gradually, at least in children’s vocabulary, from one or a few broad categories around the age of 2 years to several ones that can cover each of the universal expressions discretely around the age of 6–7 years (i.e., broad-to-differentiated pattern; for an overview, see Widen, 2013). However, the development in label use might not be an optimal indication of children’s developing perceptual sensitivity for the distinctions between facial expressions as knowledge of the labels does not mean that children know how to apply them correctly (Vicari et al., 2000; Widen, 2013). In addition, even though children from 6–7 years seem to be able to label all of the facial expressions (Widen & Russell, 2008), studies have shown considerable improvement of children after the age of 6 years in their ability to label, discriminate, categorize, and recognize them (e.g., Herba et al., 2006; Rodger et al., 2018; Thomas et al., 2007; Vicari et al., 2000; Widen, 2013). Therefore, a study design that does not require verbal responses and measures both correct and incorrect identifications of facial expressions as belonging to emotion categories might be more suitable for investigating the differentiation process by which children become increasingly sensitive for the perceptual distinctions between facial expressions. This could be established by making a certain emotion category salient to children after which they have to identify facial expressions as belonging to that category or not. If children learn to perceive each of the facial expressions as perceptually discrete and mutually exclusive from each of the others, this should be reflected in an increase in correctly perceiving each of the universal expressions as belonging to a discrete emotion category and a decrease of incorrectly perceiving expressions as belonging to that same discrete emotion category. As studies have shown that children are better in nonverbal emotion recognition tasks than verbal labeling tasks (Klimmt et al., 1983; Vicari et al., 2000; Widen & Russell, 2008) and 7-year-olds are able to label each of the universal expressions above chance level (Durand et al., 2007; Widen, 2013), it is likely that children from 6 years of age are to be able to identify them nonverbally above chance level as well.

Context-Dependent Categorical Perception of Facial Expressions

Ekman and Friesen (1971) proposed that there are six categories of facial expressions (i.e., “anger,” “disgust,” “fear,” “happy,” “sad,” and “surprise”) that are perceived in a similar way by humans across different cultures (also see Darwin, 1872; Ekman, 2003). This seems to be the reason why the abundance of the aforementioned studies on how children develop in perceiving facial expressions includes a selection of these categories. However, recent studies have shown that the sociocultural upbringing of humans largely affects how humans view these categories of facial expressions, suggesting that these six categories are not perceived similarly across different cultures (e.g., Jack et al., 2012). Sociocultural differences in categorical perception is a phenomenon that is not restricted to facial expressions as it is also evidenced in the perception of categories of color and phonemes (e.g., Barrett, 2006; Goldstone & Hendrickson, 2010; Özgen & Davies, 2002; Ozturk et al., 2013). These cultural differences likely have a social constructivist origination, meaning that the perception of categories of facial expressions and their meaning emerge from context-dependent functional relations between the perceived expressions and the possible actions that are adaptive in response to them (cf. [social] affordances; Gibson, 1979; Walker-Andrews, 1997; Widen, 2013; Withagen & Michaels, 2007; also see Traible & Pauen, 2007).

With the term “context,” we refer to perceptual information that, together with the perceived facial expression, can be meaningfully connected to potential actions of the observer. Social referencing provides an illustrative example of the ability of preverbal infants to use another person’s emotional expression to guide their own behavior in various contexts (Klimmt et al., 1983; Ruba & Repacholi, 2020; Walker-Andrews, 1997). Context includes learned social practices and norms in relation to the perceived facial expression, the specific (social) environment in which the facial expression is perceived, but also the emotional state and personal history of the observer. The idea that specific experiences shape the way facial expressions are perceived is in line with the finding that children who have experienced an abnormal amount of hostility and anger in their life tend to respond differently to angry expressions in respect to other expressions and in comparison with their peers (Ardizzi et al., 2015; Pollak et al., 2009; Pollak & Tolley-Schell, 2003). Hence, perception of facial expressions is shaped by specific exposure and experience with facial expressions over time and across specific (cultural) contexts (see Goldstone, 1998; Goldstone & Hendrickson, 2010; Kelly et al., 2011; Leppänen & Nelson, 2009; Pollak & Kistler, 2002; Pollak et al., 2009).

The extent to which a context can shape the facial expressions we perceive is also evidenced by the fact that humans are likely to perceive expressions (e.g., happiness) that are congruent with the context (e.g., a birthday party) in which they are perceived (Carroll & Russell, 1996; also see Ruba et al., 2019). As the meaning (cf. affordances) of facial expressions is intrinsically context dependent (Barrett, 2006; Carroll & Russell, 1996; also see Marchi & Newen, 2015), it might explain why children tend to perform much better in facial expression recognition tasks and labeling tasks when the task is preceded with a context in terms of a story, behavioral consequences, or pictures depicting a scene that fits to the perceived facial expressions (e.g., Vicari et al., 2000; Widen & Russell, 2004, 2013). For these reasons, it seemed appropriate to investigate the developing ability of children to identify facial expressions according to the prototypical context in which they are normally conceived in their culture. In doing so, we expected that young and relatively unexperienced children are likely to misidentify all kinds of facial expressions as belonging to the
Flexible Boundaries Between Categories of Facial Expressions

Widen and Russell (2003, 2008) found that incorrect labeling by children up to 7 years of age seems to be systematic rather than random and that children are likely to use the same label to describe different expressions with a similar conceptual valence (either pleasant or unpleasant) or level of arousal. However, this result might also be partly due to the fact that children are biased to use some labels more than others in describing the different facial expressions, favoring labels that are more commonly used than others (Widen, 2013). Moreover, when a specific facial expression such as fear is incorrectly identified as belonging to a prototypical context in which anger is normally expressed, this does not necessarily have to imply that expressions of anger are also incorrectly identified as belonging to a prototypical context in which fear is normally expressed. For example, given a certain context such as a bullying situation at school, expressions of disgust might be perceived as expressions of anger by children as both expressions could be seen as suitable to the situation (Danovitch & Bloom, 2009). However, that does not imply that they would perceive expressions of disgust as expressions of anger in another context such as a display of infested food where expressions of anger are generally less suitable. Therefore, when facial expressions need to be categorized as belonging to context that makes a particular emotion category salient, we do not expect that incorrect categorizations are always necessarily bidirectional. Such a result would stress that the differences between the facial expressions that represent different emotional states do not lie on internalized orthogonal continua such as pleasant–unpleasant or high arousal–low arousal, on which strict category boundaries can be identified (see Bimler & Kirkland, 2001; Widen & Russell, 2008; Young et al., 1997). Rather, the appearance of category boundaries might be much more flexible and contextualized, as has been shown to be the case for the speech sound categories (Case et al., 1995; Hasselman, 2015).

The Current Study

In the current study, a two-alternative forced-choice task was used in which 6- to 9-year-old children growing up in a West European culture (i.e., the Netherlands) were instructed to identify seven different facial expressions ("anger," "disgust," "fear," "happy," "sad," "surprise," and "neutral") presented in succession as belonging to a target emotion category or not in six different conditions. Each condition was preceded with a presentation of a prototypical context that was "anger," "disgust," "fear," "happy," "sad," or "surprise." To operationalize their ability to identify facial expressions into emotion categories displayed by prototypical contexts (called “perceptual sensitivity” henceforth), the results were analyzed using signal detection theory (SDT). SDT takes both correct responses (hit ratio) and incorrect responses (false alarm ratio) into account in constructing a single index of perceptual sensitivity for all categories together and for each category separately (see Stanislaw & Todorov, 1999; Thomas et al., 2007). There are several advantages of applying SDT to the measures of the proposed forced-choice task. First, it capitalizes on both correct and incorrect responses statistically, similar to the qualitative analysis on correct and incorrect labeling of facial expressions by children as Widen and Russell did (see Widen, 2013, for an overview). Second, it teases apart the response bias respondents can potentially have from the perceptual sensitivity (Durand et al., 2007; Lynn & Barrett, 2014). Response bias refers to the potential general tendency to identify facial expressions as either belonging or not belonging to the prototypical context, independent from the type of prototypical context and/or the facial expressions shown within a condition. This seems especially relevant as Widen and Russell (2003, 2008) have shown that children are biased in using some labels more frequently than others in describing facial expressions, which could mean that children in the current study might be more inclined to identify facial expressions as belonging to a certain category than others. Third, this operationalization of perceptual sensitivity has considerable conceptual overlap with the notion of “differentiation” as put forward by Gibson and Gibson (1955), as the sensitivity for perceiving distinctions leading to differential responses plays a central role in both. Finally, Shannon’s Diversity Index (see Begon et al., 1996; De Jonge-Hoeckstra et al., 2020) was used to index the diversity of incorrect responses (i.e., kinds of false alarms) across age groups.

Method

Participants

The final sample consisted of 164 primary school children (98 girls) between ages 5 and 9 (M_age = 7 years and 7 months, SD = 1 year and 2 months) from four different school grades from five different schools in the Netherlands (see Table 1 for an overview).

Three participants (two in second grade and one in fifth grade) were excluded from the analysis because of computer failure. The gender of two participants (one in second grade and one in fifth grade) was registered incorrectly and is therefore missing in the analysis. In constructing age groups, we aimed to minimize intra-group variation and maximize intergroup variation in terms of the (social) contexts in which they normally view facial expressions. We reasoned that children within a school grade are more likely to be similar in the types of contexts in which they perceive facial expressions and display more commonalities in their ability to interact within these contexts than children of the same chronological age per se. Therefore, the age groups of children comprised children attending the same school grade and hence have some overlap, although very little in terms of chronological age. Prior to participation, informed consent was acquired from one parent of each child. This study was part of a research program titled “The Perceptual Basis for Emotion Recognition and Perspective Taking” and was approved by the ethical committee of the Faculty...
of Social Sciences, University Utrecht under protocol number FETC18-037.

Materials

The Facial Expression Perception (FEP) task consisted of 72 trials. In each trial, a photograph was presented with a face displaying one of the six facial expressions (i.e., anger, disgust, fear, happy, sad, or surprise) or the neutral expression. The person displaying the facial expression could be either male or female and was either a child, middle-aged adult, or an older adult. The photographs were shown in grayscale with the background and hair around the face removed to reduce distraction and to ensure the saliency of the facial expressions.1 Trials were divided over six conditions. Each condition contained six trials displaying one of the expressions that fit the emotion category of that condition and six trials displaying the remaining expressions and the neutral expression as distractor emotions. Within each condition, both the target emotion and distractor emotions were each expressed three times by a male face and three times by a female face. For each gender, two were children, two were middle-aged people, and two were older people. Trials were shown on a 15-in. display with a resolution of 1,280 × 800 pixels, and E-prime 2 was used for stimuli presentation and data collection (Schneider et al., 2012). Participants responded by using the computer keys “L” (marked blue) for whenever they perceived the target emotion and “D” (marked red) for whenever they did not perceive the target emotion.

Procedure

Warming Up and Training

The task took place in a quiet room in the school the children were attending. After being seated in the room, the experimenter would typically chat with the child and asked whether they would like to play a small game about emotions. Next, the experimenter explained that she would sit next to the child during the task and that the child could always take a break, stop the game, or ask questions if needed and, finally, that there are no wrong answers.

Prior to the task, children did a short training session to get accustomed to pressing the computer keys and specifically to press the blue key to confirm the presence of a target and the red key otherwise. To this end, they were instructed to push the blue key whenever a blue screen was presented and otherwise press red. On four different trials, participants were shown successively a blue screen and a red screen twice and could only proceed to the next trial if they pressed the correct button. Although a few children made an error on the first training trial, all children seemed to understand the task after and responded correctly on the last three trials. After the training session, the experimenter asked if the child had any further questions or required a bathroom break, and if not, she would proceed to the first condition of the FEP task.

Testing Phase

Each condition started with a presentation of a prototypical context for a particular emotion whereafter children engaged in a forced-choice procedure in which they identified facial expressions as either belonging to that prototypical context or not. At the start of each condition of the FEP task, emotions were briefly discussed according to a script by describing situations with visual aids in which each of the six emotions would be appropriate. With each emotion, an example was given of a situation, typically leading to experiencing that particular emotion. This created the necessary context for the subsequent task in which children had to identify facial expressions as belonging to that context or not.

For happiness, pictures of an ice cream, a playground, and jumping people on the beach were shown and the child was told that most people are happy when they get to eat ice cream.

For anger, pictures showed two children, with one child getting a considerably bigger piece of cake than the other; a mother disciplining her child; a thief running away with a large sum of cash; and an almost intact ice cream in an upside-down position on the street. The child was told that a child could

### Table 1

| Sample Characteristics in Terms of Age in Years (Mean and Standard Deviation), Distribution of Children Across Gender, and Distribution of Children Across School Grades and Their Accompanying Age Groups |
|---------------------------------|----------------|----------------|----------------|
| **Age group**                   | **School grade** | **Boys** | **Girls** | **Total** |
| Six-year-olds                   | 2              | 16       | 28        | 45        |
| Seven-year-olds                 | 3              | 16       | 17        | 33        |
| Eight-year-olds                 | 4              | 13       | 24        | 37        |
| Nine-year-olds                  | 5              | 19       | 29        | 49        |
| **N**                           | **M** | **SD** | **M** | **SD** | **M** | **SD** |
| Boys                            | 6.30  | 0.50  | 6.14  | 0.32  | 6.20  | 0.40  |
| Girls                           | 6.94  | 0.32  | 6.91  | 0.36  | 6.92  | 0.33  |
| Total                           | 8.11  | 0.38  | 8.10  | 0.35  | 8.11  | 0.36  |
| Total                           | 8.86  | 0.29  | 9.01  | 0.35  | 8.95  | 0.33  |

1. Pictures were retrieved from the Internet using Google Search, using the search word “expression” in combination with each of the tested expressions (e.g., “sad expression”) and filtered on “free to use, share of modify.” A pilot study was performed to test the validity of the stimulus material. Students of the faculty of Social Sciences of Utrecht University (N = 27, 20 were female, M_age = 20 years and 4 months, SD = 1 year and 6 months) participated on voluntary basis with informed consent. During the end of a lecture, students completed an online expression labeling task via their smartphone in silence. Seventy-two pictures of expressions were presented in succession and in random order. Participants labeled each expression by pressing one out of seven options: the six used expressions and the neutral expression. Nine expressions were labeled only 50% or less correct and were therefore replaced. Two independent raters labeled all replacements correctly. With the remaining pictures, students labeled 83.86% correct on average, which is typical performance level for adults (cf. Lawrence et al., 2015) and, in our view, indicated that the material is valid for experimental use (see online supplemental materials for the pilot data).
get angry if someone else would intentionally break his or her new toy.

For sadness, pictures were shown of a child in front of a window and looking at the rainy and cold weather outside, a broken smartphone, a funeral scenario, and devastating-looking football players after losing a match. The child was told that people tend to get very sad when their pet passes away.

For fear, pictures were shown of someone wearing a blindfold and about to receive a cake, a nicely packed present, someone waiting around a corner to surprise another person who is approaching that corner, and a car that is parked on a street that is curved upside down. The child was told that people would be surprised if it would snow in the middle of the summer.

For disgust, pictures of dog poo on the street, a dead and half-eaten insect, a garbage can with the stench visualized as smoke/fog, and a plate of something that could once have been food were presented. The child was told that some people tend to get disgusted and say “yucky” (“bah” in Dutch) whenever they have to eat something they do not like to eat, such as “brussels sprouts” (for Dutch children, brussels sprouts are notorious for their bitter taste).²

After each emotion was briefly described, the experimenter asked children to describe a situation that would result in experiencing the discussed emotion. Answers were mildly praised. Before the FEP task started, the experimenter explained that a series of faces would be shown in succession, each expressing an emotion, and that they had to press the blue button whenever the face was expressing the same emotion as they just discussed, and otherwise the red button. After this instruction in the first condition, the experimenter would ask which button had to be pressed upon seeing the target emotion. Children would typically respond by saying “the blue button,” and if not (which happened only in a few cases), the experimenter would repeat the instruction, whereafter they typically gave the correct answer, revealing that they understood the task instructions.

During the presentation of the trials in each condition, children were wearing headphones in order to reduce possible distracting noises from the surrounding area and to reduce the tendency to talk during the task. The order of presentation of the conditions and pictures within the conditions was randomized. After children had completed the six conditions, they were asked if they liked to play the game and were given a sticker to thank them for participating, after which they returned to their classroom.

Measures and Coding Data

Perceptual Sensitivity

On each trial, reaction time (RT; in milliseconds) and performance (correct or incorrect) were registered. Performance of participants on each trial was recoded into a “hit” whenever participants correctly identified the perceived facial expression as belonging to the emotion category, “miss” whenever they failed to identify the perceived facial expression as belonging to the emotion category, “false alarm” when they incorrectly identified the perceiving facial expression as belonging to the emotion category, or “correct rejection” when they correctly did not identify the perceived facial expression as belonging to the emotion category. Next, perceptual sensitivity (in SDT, known as “d-prime”) was calculated for each participant for the whole task and per condition, using the following formula: \( d' = z^{-1}(HIT) - z^{-1}(FA) \), where FA is the proportion of false alarms on the trials that contained a distractor, HIT is the proportion of hits on the trials that contained a target, and \( z \) is the inverse of the normal distribution function (MacMillan & Creelman, 2004; for an extended overview, see Stanislaw & Todorov, 1999). A simple way to interpret the perceptual sensitivity index is that a score above 0 means that the proportion of hits is larger than the proportion of false alarms, meaning that the participant is able to distinguish the target emotion(s) from the distractors above chance level (MacMillan & Creelman, 2004). To avoid divisions by zero, the proportion of false alarms and proportion of hits were adjusted whenever they were either 1 or 0 by replacing the 0s with \( X_{adj} = \frac{1}{2} \) and the 1s by \( X_{adj} = 1 - \frac{1}{2} \) before \( d' \) was calculated, as suggested by MacMillan and Creelman (2004).

Criterion Bias Index

The criterion bias index signifies the extent to which participants are biased to either confirm or disconfirm perceiving the target, regardless of the expression shown, and was used as a measure of the reliability of the data (see Stanislaw & Todorov, 1999). The criterion bias index (c) was calculated for each participant. In order to calculate c, the following formula was used:

\[
   c = \frac{(c^{-1}(HIT) + c^{-1}(FA))}{2}
\]

Negative scores indicate a tendency to confirm the presence of a target, positive scores indicate a tendency to disconfirm the presence of a target, and scores around 0 indicate that the participant was not biased in any direction (Stanislaw & Todorov, 1999). In order to calculate c whenever the proportion of false alarms and proportion of hits were either 1 or 0, these proportions were adjusted using the same formulas as presented in the section above.

Shannon’s Diversity Index of False Alarms

Shannon’s Diversity Index (H) was used to index the diversity of distractor facial expressions that were incorrectly identified (i.e., false alarms) as belonging to the prototypical contexts of each target emotion for participants of each age group. This index can reveal to what extent the type of false alarms in each condition made by participants of an age group is either random or nonrandom. If it is random, it means that participants of an age group made many different kinds of false alarms in a condition. For instance, 5-year-olds might mistakenly identify expressions of anger, fear, sadness, and surprise as belonging to a prototypical context for disgust. Conversely, if it is nonrandom, it means that participants of an age group and for a target emotion made one or a few specific kinds of false alarms. For example, 9-year-olds might mistakenly identify an expression of anger as belonging to a

² Three independent raters categorized the prototypical contexts, consisting of the collection of pictures, in terms of one out of six emotion categories used in the experiment. All raters categorized the prototypical contexts as belonging to their emotion categories as intended (100% agreement), revealing that the found effect was unlikely to be due to ambiguity in terms of how the prototypical contexts were constructed.
prototypical context for disgust but do not make other mistakes in identifying expressions as belonging to the prototypical context of disgust. We calculated $H$ for each age group and for each condition using the following formula: $H = - \sum_{i=1}^{s} p_i \ln p_i$, where $s$ is the number of different kinds of possible false alarms that were made in each condition (which can be between one and six as there are six possible kinds of false alarms), and $p_i$ is the proportion of observed false alarms of a particular kind relative to the total number of false alarms across all kinds of false alarms that have been made by children of that age group in a particular condition (see Begon et al., 1996). As $p_i$ is a proportional measure, $H$ is invariant with respect to the absolute number of false alarms that a particular age group made in a particular condition. In other words, $H$ reflects the distribution of false alarms independently from the absolute number of false alarms made by an age group or in a condition. When $H$ is high, it means that the proportions of false alarms are uniformly and randomly distributed across the different distractor emotions, indicating that participants of that age group identified a high diversity of facial expressions as belonging to the context that described the target emotion. Conversely, a low $H$ indicates that the proportions of false alarms are distributed unevenly across the different distractor emotions, indicating that participants of that age group identified a low diversity of facial expressions to the context that described the target emotion. The average and standard deviation of $H$ across the six conditions were calculated for each of the four age groups. Finally, the calculated values of $H$ were transformed into Shannon’s equitability ($E_H$) to fit into a scale that runs from 0 (no diversity and no randomness) to 1 (maximum diversity, complete randomness) in order to easily interpret the differences in values across age groups and emotion categories. This transformation was performed, using the following formula: $E_H = H / (\frac{H}{p_{\text{max}}})$.

**Data Preparation and Analysis**

**Analysis of RT**

RTs (in ms) were averaged across trials and across conditions for each participant. In order to test whether average RTs differed across age group, a one-way analysis of variance (ANOVA) was performed using the average RT per participant. All assumptions for performing the ANOVA were checked. Two participants were removed from the analysis, as their average RT deviated 3 SD from the average RT in their age group. According to Levene’s test, variances in RTs between children from different age groups were not equal, $F(3, 158) = 7.89, p < .001$. Data transformations as suggested by Howell (2007) and Tabachnick and Fidell (2007) did not seem to change this outcome. Therefore, the alpha was set to .01 in order to reduce the chance on a Type I error (Moder, 2010), and the nonparametric Games–Howell post hoc test was used to analyze which age groups differed from each other in terms of average RTs (Ruxton & Beauchamp, 2008). Effect sizes for the significant group differences in RT were calculated in terms of Cohen’s $d$.

**Analysis of SDT Measures and Shannon’s Diversity Index of False Alarms**

In order to test if the children in our sample were able to relate each of the facial expressions to its contextual description above chance level, we tested whether the perceptual sensitivity ($d’$) was above zero for each of the four age groups and for each of the six conditions, using 24 separate $t$ tests. In order to account for an inflated Type I error, a Bonferroni correction was applied, meaning that $\alpha_{\text{critical}}$ of .05 was divided by 24, resulting in $\alpha_{\text{altered}}$ of .002. All assumptions for performing these tests were met. In order to test whether average perceptual sensitivity ($d’$) for the different expressions combined on the FEP task differed across participants from the four age groups, a one-way ANOVA with a Bonferroni post hoc tests was performed. All assumptions for performing the ANOVA were checked and met, and no outliers were removed as all obtained sensitivity scores did not deviate more than 3 SD from the average. Effect sizes for the significant group differences (as indicated by the Bonferroni comparison test) in average $d’$ on the FEP task were calculated in terms of Cohen’s $d$. To test whether children across different age groups differed on $d’$ on each of the six emotion categories while taking the within-subjects variance across the conditions into account, a multivariate ANOVA (MANOVA) was performed with Bonferroni post hoc comparison tests, using age group and gender as the independent variables and the $d’$ for each emotion category as the dependent variables. Pillai’s trace was used as an indication of whether age groups generally differed from each other, whereas Bonferroni post hoc comparison tests were used to indicate age group differences for each emotion category. Effect sizes for the significant group differences as indicated by the Bonferroni test in $d’$ for each emotion category were calculated in terms of Cohen’s $d$.

Next, $\eta^2$ of the interaction of emotion category and age group of all the motion categories were calculated in order to compare the sizes of age difference in $d’$ between the emotion categories. The $\eta^2$ of the intercept of all the motion categories was calculated in order to compare the $d’$ of participants for each emotion category independent from age group. Three univariate outliers were removed as the obtained $d’$ deviated more than 3 SD from the average $d’$ in that age group. In addition, one multivariate outlier was removed, based on a deviation from the expected the Mahalanobis distance ($p < .001$ $\chi^2$ distribution). All assumptions for performing the MANOVA were checked and met.

In order to test whether the criterion bias index $c$ for the FEP task differed across participants from different age groups, a one-way ANOVA was performed using the $c$ for the FEP task per participant. All assumptions for performing the ANOVA were checked and met. An additional MANOVA with post hoc Bonferroni comparison test was performed in order to check whether children across age groups differed in $c$ for each emotion category separately. Pillai’s trace was used as an indication of whether age groups generally differed from each other, whereas Bonferroni post hoc comparison tests were used to indicate age group differences for each emotion category. Effect sizes for the significant group differences as indicated by the Bonferroni comparison test in $d’$ for each emotion category were calculated in terms of Cohen’s $d$. Four univariate outliers were removed as the obtained $c$ deviated more than 3 SD from the average $c$ in that age group. All assumptions for performing the MANOVA were checked and met.
To test whether children from different age groups differed on H, a one-way ANOVA with Tukey’s honestly significant difference (HSD) post hoc tests was used. Effect sizes for the significant group differences (as indicated by the Tukey’s honestly significant difference (HSD) tests) in H were calculated in terms of Cohen’s d. All assumptions for performing these analyses were checked and met. Finally, a difference score in H was calculated between the 9-year-olds and 6-year-olds in order to observe age trends in this measure.

Results

RT

Average RT on the task seemed to differ across children from different age groups, \( F(3, 158) = 18.33, p < .001, \eta^2 = .26 \). Non-parametrical post hoc Games–Howell tests indicated that 6-year-olds (\( M = 3,438 \) ms, \( SD = 1,197 \) ms, \( N = 45 \)) were significantly slower than 7-year-olds (\( M = 2,552 \) ms, \( SD = 612 \) ms, \( N = 33 \), \( p < .001 \), Cohen’s \( d = 1.04 \)), 8-year-olds (\( M = 2,344 \), \( SD = 653 \), \( N = 37 \), \( p < .001 \), Cohen’s \( d = 1.13 \)), and 9-year-olds (\( M = 2,282 \), \( SD = 640 \), \( N = 49 \), \( p < .001 \), Cohen’s \( d = 1.20 \)). No other age group differences were found (all \( p > .05 \)).

Perceptual Sensitivity and Response Bias Measures

The distribution of \( d' \) scores for each emotion category of all children differed significantly from 0 (all \( t \) tests, \( p < .001 \); see online supplemental materials), indicating that on average, children had a higher hit ratio than a false alarm ratio, meaning their performance on the FEP task was above chance level. Children differed in terms of overall perceptual sensitivity (\( d' \)) across age on the FEP task, \( F(3, 160) = 16.22, p < .001, \eta^2 = .23 \). Post hoc Bonferroni comparison tests revealed that in terms of perceptual sensitivity scores, 6-year-olds (\( M = 1.44 \), \( SD = .68 \)) scored lower than 8-year-olds (\( M = 2.13 \), \( SD = .65 \), \( p < .001 \), Cohen’s \( d = 1.04 \)) and 9-year-olds (\( M = 2.25 \), \( SD = .50 \), \( p < .001 \), Cohen’s \( d = 1.36 \)), and 7-year-olds (\( M = 1.77 \), \( SD = .62 \)) scored lower than 9-year-olds (\( p = .003 \), Cohen’s \( d = .85 \)). No other age group differences were found (all \( p > .05 \); see online supplemental materials). These group differences suggest a positive relation between age and overall perceptual sensitivity (see Figure 1). A MANOVA yielded a similar result in showing that children across age generally differed in perceptual sensitivity for the individual emotion categories on the FEP task, Pillai’s trace = .28, \( F(18, 459) = 2.59, p < .001, \eta^2 = .092 \). Post hoc Bonferroni comparison tests results suggest a general positive relation between age and \( d' \) for each emotion category (see Figure 2 and online supplemental materials). In addition, children across the tested age groups differed in their perceptual sensitivity for the emotion categories in the following order of size of difference (from large to small): “surprise” (\( \eta^2 = .84 \)), “sad” (\( \eta^2 = .68 \)), “disgust” (\( \eta^2 = .44 \)), “angry” (\( \eta^2 = .33 \)), “happy” (\( \eta^2 = .11 \)), “fear” (\( \eta^2 = .09 \)), “sad” (\( \eta^2 = .07 \)), “disgust” (\( \eta^2 = .06 \)), “angry” (\( \eta^2 = .05 \)), “happy” (\( \eta^2 = .04 \)), and “surprise” (\( \eta^2 = .03 \)). Based on the \( \eta^2 \) of the intercepts of the individual emotion categories that control for age, perceptual sensitivity for the emotion categories had the following order (from high to low): “happy” (\( \eta^2 = .94 \)), “angry” (\( \eta^2 = .84 \)), “sad” (\( \eta^2 = .77 \)), “surprise” (\( \eta^2 = .76 \)), “disgust” (\( \eta^2 = .76 \)), and “fear” (\( \eta^2 = .68 \)). Average response bias index did not differ across age, \( F(3, 160) = 1.94, p = .126 \), and the average response bias index for the entire test population was close to 0 (\( M = -.004 \), \( SD = .33 \)). However, a MANOVA revealed a difference in \( c \) between children across age for the individual emotion categories combined per age group (see Table 1).
categories on the FEP task, Pillai’s trace = .15, $F(18, 459) = 1.35$, $p = .050$, $\eta^2 = .050$. Post hoc Bonferroni comparison tests revealed only one age group difference, namely, that 6-year-olds ($M = .26$, $SD = .44$) less often confirmed the presence of the emotion category than 9-year-olds ($M = .04$, $SD = .28$, $p = .028$, Cohen’s $d = .59$; see online supplemental materials for all Bonferroni comparison results).

**Shannon’s Diversity Index of False Alarm Distractors**

Children across age groups differed in average $H$, $F(3, 160) = 19.76$, $p < .0001$; $\eta^2 = .27$. Post hoc Tukey’s HSD test results suggest a general negative relation between age and $H$ as they indicated that 6-year-olds ($M = 1.64$, $SD = .16$) scored higher than 7-year-olds ($M = 1.53$, $SD = .12$, $p = .01$, Cohen’s $d = .78$), 8-year-olds ($M = 1.49$, $SD = .13$, $p < .0001$; Cohen’s $d = 1.03$), and 9-year-olds ($M = 1.40$, $SD = .17$, $p < .0001$; Cohen’s $d = 1.45$); 7-year-olds scored higher than 9-year-olds ($p = .001$, Cohen’s $d = .88$) but not higher than 8-year-olds ($p = .69$); and 8-year-olds scored higher than 9-year-olds ($p = .04$, Cohen’s $d = .59$). The frequencies of false alarms per age group and per target emotion can be found in the online supplemental materials. The standardized scores of $H$ ($E_{iH}$) are presented in Table 2 and show that young children have larger values of $E_{iH}$ than older children for most expressions, meaning that the distribution of false alarms is more uniform for younger children than older children for these expressions. This is further illustrated in Figure 3, which shows that distributions of proportions of false alarms for most emotion categories seem to be rather uniformly distributed across the different distractor expressions for 6-year-olds compared to older children, especially 9-year-olds. The largest decline in $E_{iH}$ across age was for the expression “surprise,” followed by “fear” and “disgust,” suggesting that children develop in distinguishing these particular expressions from others across the tested age range (see Table 2). The distribution of false alarms does not seem to change for the emotion categories “sad” and “happy” across ages, considering the small difference scores of $E_{iH}$ between 6-year-olds and 9-year-olds (see Table 2).

**Table 2**

<table>
<thead>
<tr>
<th>Age group</th>
<th>Anger ($E_{iH}$)</th>
<th>Disgust ($E_{iH}$)</th>
<th>Fear ($E_{iH}$)</th>
<th>Happy ($E_{iH}$)</th>
<th>Sad ($E_{iH}$)</th>
<th>Surprise ($E_{iH}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>.95</td>
<td>.93</td>
<td>.95</td>
<td>.81</td>
<td>.94</td>
<td>.97</td>
</tr>
<tr>
<td>7</td>
<td>.88</td>
<td>.76</td>
<td>.86</td>
<td>.80</td>
<td>.94</td>
<td>.89</td>
</tr>
<tr>
<td>8</td>
<td>.92</td>
<td>.82</td>
<td>.73</td>
<td>.78</td>
<td>.92</td>
<td>.80</td>
</tr>
<tr>
<td>9</td>
<td>.85</td>
<td>.72</td>
<td>.86</td>
<td>.78</td>
<td>.92</td>
<td>.72</td>
</tr>
<tr>
<td>Difference score</td>
<td>.10</td>
<td>.21</td>
<td>.09</td>
<td>.03</td>
<td>.02</td>
<td>.25</td>
</tr>
</tbody>
</table>
In the current study, we investigated the difference in ability across 6- to 9-year-olds to identify emotional facial expressions as belonging to prototypical contexts for each emotion. We found that the perceptual sensitivity of children of all age groups was above chance level for all emotions and that older children are more perceptually sensitive to perceive each of the six expressions as distinctly belonging to specific emotion categories than younger children (cf. Herba et al., 2006; Rodger et al., 2018; Thomas et al., 2007). Specifically, with increasing age and experience, children become better in correctly identifying facial expressions as belonging to their prototypical contexts and made fewer mistakes in terms of incorrectly identifying facial expressions as belonging

**Figure 3**
Proportions of False Alarms Distributed Across Distractor Expressions per Emotion Category for Each Age Group

Note. The dotted vertical reference line (on 16.66) indicates the proportions of false alarms per emotion category that would be expected in an uniformal distribution. False alarm proportions deviating from this line indicate a low Shannon’s Diversity Index for the emotion category to which the false alarm proportions belong. The average number of false alarms per emotion category decreases as the age of children is increasing. See the online article for the color version of this figure.
to the prototypical contexts. The same pattern was observed when looking at the age-group differences in terms of the perceptual sensitivity for each of the used expressions separately, although these differences were more pronounced for some emotion categories compared to others. Specifically, perceptual sensitivity for the facial expressions “surprise” “sad,” and “disgust” seemed to develop stronger across the tested age groups than the perceptual sensitivity for the facial expressions “fear,” “happy,” and “angry.” A similar age-related pattern has been observed in other studies that have shown that children within this age range improve in identifying these facial expressions (cf. Durand et al., 2007; Gao & Maurer, 2010; Rodger et al., 2015). In addition, the perceptual sensitivity for identifying facial expressions as belonging to their prototypical contexts across all age groups had the following order (from high to low): “happy,” “angry,” “sad,” “surprise,” “disgust,” and “fear” (cf. Durand et al., 2007; Herba et al., 2006; Rodger et al., 2018; Segal et al., 2019; Thomas et al., 2007; Vicari et al., 2000; Widen & Russell, 2013). The fact that the perceptual sensitivity for happy facial expressions was higher than for any other expression is in line with studies that have shown that humans of all ages tend to be very accurate in classifying, discriminating, and identifying happy faces (Gao & Maurer, 2009; 2010; Herba et al., 2006; Rodger et al., 2018; Thomas et al., 2007; Vicari et al., 2000; Widen, 2013). Perhaps, this finding could be explained by the idea that the tested children have a disproportionate amount of experience with perceiving this particular expression as they are generally approached in a positive manner in the Netherlands (see UNICEF, 2007). In addition, perhaps the pleasant atmosphere experimenters generally try to maintain during testing might unintentionally cause children to have an increased perceptual sensitivity for happy expressions like in the current study. More generally, the found age differences suggest that the children up to 9 years develop in their ability to perceive facial expressions as categorically distinct. However, this ability seems to develop further during adolescence and adulthood (see e.g., De Sonneville et al., 2002; Kessels et al., 2014; Lawrence et al., 2015). Finally, we calculated the response bias index in order to control for the possibility that children were biased to systematically confirm or disconfirm that facial expressions belonged to a certain emotion category. This bias could occur as an effect of fatigue, lack of concentration, or not understanding a certain emotion category or the instructions of the FEP task. The average response bias index was close to 0 for all children and did not differ across ages, indicating that children understood the task specifics and the emotion categories as displayed by the prototypical contexts.

Next, we found a negative relation between age group and the diversity of false alarms, meaning that older children identify fewer kinds of facial expressions as belonging to an emotion category than younger children (cf. Widen & Russell, 2008). For example, 6-year-olds incorrectly identified a high diversity of expressions as belonging to the prototypical context of surprise such as happy, anger, sadness, fear, neutral, and disgust. However, 9-year-olds would more correctly identify expressions of fear as belonging to surprise but hardly any other expressions. The age-related decline of diversity of false alarms for each target facial expression is most pronounced for expressions of anger, disgust, and surprise, which is in accordance with other study results using various methods in showing that children between 6 and 9 years of age develop vastly in terms of perceiving these facial expressions (cf. Gao & Maurer, 2009, 2010; Herba et al., 2006; Rodger et al., 2018; Thomas et al., 2007). Together, these results show that with age and experience, children seem to improve in differentiating facial expressions, allowing them to better identify them as belonging to their respective prototypical contexts.

Finally, the kinds of misperceptions for each target emotion were qualitatively analyzed in order to assess if incorrect identifications across target emotions were made bidirectionally. We found that when children incorrectly identified a certain distractor facial expression such as anger for a target emotion such as disgust, this did not imply that the opposite was also true (cf. Widen & Russell, 2013). That is, if anger was the target emotion, children would not necessarily incorrectly identify disgust as belonging to that emotion but would more often incorrectly identify neutral or fear as anger. Indeed, it is conceivable that a prototypical context for the emotion disgust (e.g., someone placing a plate of rotting food on your table) could evoke feelings or displays of anger toward the cause of that context. Alternatively, a prototypical context of anger (e.g., someone breaking your brand-new toy) might be less likely to evoke feelings of disgust (cf. Danovitch & Bloom, 2009). Sad expressions were often incorrectly identified as happy expressions, while angry expressions and expressions of disgust were most often incorrectly identified as sad expressions. The fact that sad expressions were often incorrectly identified as belonging to the happy category seems peculiar and is not in line with the findings of Widen and Russell (2008), who have performed a similar analysis on mislabeling facial expressions by children up to 5 years of age. Therefore, we can merely speculate in accounting for this finding. For example, perhaps children found the sad face to be suitable for the prototypical context of happy that included the statement that eating ice cream makes people happy as they did not have an ice cream at the time of testing, making them feel sad about it and leading them to respond accordingly. Only the combination of expressions of fear and surprise seemed to operate bidirectionally in that expressions of fear were often incorrectly identified as expressions of surprise and vice versa (cf. Widen & Russell, 2008). These findings contradict the idea that especially expressions that are dissimilar in emotional valence (i.e., happy is pleasant, sad is not) are not likely to be confused with each other. Instead, these results imply that emotional states cannot be placed on simple continua that separate the full domain of human emotions into distinct nonoverlapping categories along a dimension such as either pleasant or unpleasant.

In the current study, children had a West European background that could be more generally classified as a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) background (Krys et al., 2016). As facial expressions have different functions and meanings for humans depending on their background, response bias and diversity of identifications are likely to differ across populations with a different cultural background (Jack et al., 2012; Krys et al., 2016; Smith et al., 2013). This means that the found age differences in these measures in the current study cannot be generalized to children with a different cultural background. It would be interesting to conduct this research across humans with differing cultural backgrounds that might elucidate the age-related differences in cultural-specific functional relations between the perception of facial expressions as they occur in (social) contexts and the adaptive responses to them.

3 Apart from a small and, in our view, negligible difference between 6-year-olds and 9-year-olds in the happy emotion category, with 6-year-olds being slightly more conservative than 9-year-olds.
Next, we found a negative relation between age group and average RTs on the FEP task, meaning that as children grow older, they need less time to identify facial expressions as belonging to their emotion categories. Gao and Maurer (2009, 2010) have shown that in comparison to 5-year-old children, children 9 and 10 years of age also need less time to identify facial expressions and are better able to identify facial expressions that are presented with low intensities such as a mild expression of anger (for comparable results and methods, see Rodger et al., 2018; Thomas et al., 2007). Together, their results point to the idea that as children grow older, they require less visual information to identify facial expressions and identify them as belonging to distinct emotion categories. From a perceptual learning account, these results could be explained by the idea that with experience, children become more perceptually and selectively attuned (i.e., sensitive) to the invariant and distinctive features of each facial expression and the context in which they are presented (see Gibson, 1979; Goldstone, 1998; Hellendoorn et al., 2015; Kellman & Massey, 2013; Leppänen & Nelson, 2009; Pollak & Kistler, 2002; Pollak et al., 2009; Zebrowitz, 2011). This form of learning is referred to as “differentiation” (Adolph & Kretch, 2015; Genesee et al., 1995; Gibson & Gibson, 1955; Gibson & Pick, 2000; Goldstone, 1998; Goldstone & Styvers, 2001; also see Honey et al., 2010, for a comparable notion of differentiation called “acquired distinctiveness”). However, the age-related decrease in RT found in the current study and others could also be due to other factors such as an enhanced (oculo-) motor control. Therefore, although the idea that increases in perceptual sensitivity for the categories of facial expressions are due to differentiation seems plausible in our view, it remains speculative based on the current data and should be investigated further in future studies, using, for instance, eye-tracking technology.

Finally, we restricted the FEP task in terms of duration as to prevent effects of fatigue for the youngest age groups and in order to comply with the regulations of the ethical board of the university concerning the task demands for young children. In effect, each condition had 12 trials, which were too little data for reliable computations of Shannon’s Diversity Index of the incorrect identifications on the individual level and on the level of conditions. Therefore, Shannon’s Diversity Index was only computed on the group level. With older participants or at least without the two youngest age groups, it would be more feasible to include more trials per condition, allowing to compute Shannon’s Diversity Index of the incorrect identifications on the individual level. This could be used to investigate factors related to individual differences in Shannon’s Diversity Index of the incorrect identifications. For instance, as recent literature suggests that children with autistic spectrum disorder have trouble in perceptually differentiating (social) information (Hellendoorn et al., 2015), this deficiency is likely to be evidenced in a rather high diversity of incorrect identifications compared to their normally developing peers. Extending the FEP task in terms of adding more trials and hence the duration of the task might affect age-related memory demands of the task that needs to be controlled for.

References