Developmental Changes in Children's Categorization of Facial Cues of Emotion

Anonymous CogSci submission

Abstract

How do children learn to categorize the facial configurations classically believed to represent basic emotions? Many studies have examined when children are able to perceptually discriminate between emotional facial expressions and when children are able to verbally label these expressions. However, while these studies provide important information about the timeline of emotional development, they give less information about the nature of children's category representations for different facial configurations. For instance, emotion concepts may emerge from children's perceptions of facial configurations along the dimensions of valence and arousal. To evaluate how 3- to 7-year-old children categorize emotion concepts, we had them sort facial configurations on a grid based on whether the people were feeling "the same kind of thing". We found that while both children and adults consistently sorted faces according to the dimensions of valence and arousal, sorting faces using discrete emotion categories emerged only gradually across development, with children not demonstrating consistent use of emotion categories until approximately 5 years of age.

Keywords: face processing; emotion categories; free sorting; development

Introduction

How do children learn to categorize different facial cues that are believed to represent various emotions? While young infants do not reliably differentiate facial cues of emotion in adults, this lack of differentiation quickly changes. By around 7 months of age, children can discriminate between expressions of anger, fear, happiness, sadness, and surprise (Grossmann, 2010). However, children take much longer to learn how to map these expressions onto labels. For instance, 2-year-olds were found to use categories like "happy" and "anger" more broadly, such that they might categorize surprised, fearful and happy faces as "happy", and sad, disgusted and angry faces as "angry" (Widen & Russell, 2008). Children's ability to accurately and reliably label various facial cues of emotion continues to improve into adolescence (Montirosso et al., 2010). However, our knowledge of children's development in this area is limited by the methods used. These studies often only test children's ability to visually discriminate between different cues and to label those cues.

Our understanding of the dimensions that underlie children's representations of facial cues is also limited by these methods. Children's representations are hypothesized to be driven by the dimensions of valence (pleasant to unpleasant) and arousal (activation to deactivation; Russell, 2003). While valence is often treated as a one-dimensional measure, representing positivity and negativity in a two-dimensional space has been found to more accurately capture feelings of ambivalence and indifference (see the evaluative space grid; Larsen, Norris, McGraw, Hawkley, & Cacioppo, 2009). Traditional valence scales ask participants to rate stimuli on a scale from positive to negative. Scores in the middle of the scale could indicate that the individual feels neither positive nor negative about the stimulus (indifference), or that the individual feels equal amounts of positivity and negativity (ambivalence). Thus, allowing for positivity and negativity to exist as separate dimensions may better capture emotional representations.

Alternatively, children's representations of facial cues may be clustered by basic emotion categories (e.g., sad, happy, fear, anger, disgust, and surprise; Ekman, 1992). These two possibilities are not entirely incompatible, as children may initially perceive faces more continuously along dimensions like valence, but become more categorical and clustered in their perceptions of faces as their linguistic knowledge increases.

In the present study, we aimed to gather information about children's representations of facial cues in a more open-ended task that does not prescribe the use of specific dimensions or of pre-existing emotion labels. In the task, children judge the similarity of different expressions using the spatial arrangement method (SpAM; Goldstone, 1994). In the SpAM task, children sort images on a grid according to how similar they are to one another. Images placed closer together are assumed to represent more semantically related items. When children put images far apart, the task assumes that children view those images as less semantically related. This task is valuable because it does not limit children to prescribed categories and it does not rely on children's emotional vocabulary knowledge. The SpAM task has been found to successfully capture changes in children's semantic knowledge organization in other domains, such as plants and animals (Unger & Fisher, 2019; Unger, Fisher, Nugent, Ventura & MacLellan, 2016).

Experiment

We tested children between 3 and 7 years of age and compared their sorting behavior to that of a group of adult participants on two sets of images from the IASLab Face Set. In addition to children's sorting behavior, we also collected information on the properties of the facial configurations, including ratings of valence, arousal, positivity and negativity. We predicted that (1) the use of core dimensions such as valence and arousal in organizing facial configurations would be used consistently both by children and adults, (2) separating valence into two dimensions of positivity and negativity would have greater explanatory power, and (3) the use of "basic" emotion categories in organizing facial configurations would emerge more gradually across development.

Method

Participants

We recruited 90 children (ages 3-7; mean age = 4.94; 51 female) and 40 adults (18-21; mean age = 18.83; 30 female). 1 child was excluded as they only completed the practice phase. Data collection is still ongoing, with a target of (at minimum) 20 children in each of 4 total age groups (3-4 year olds, 4-5 year olds, 5-6 year olds, and 6-7 year olds).

Stimuli

Facial configuration stimuli. The facial configuration stimuli were selected from the Interdisciplinary Affective Science Laboratory (IASLab) Facial Stimuli Set (Gendron, Lindquist & Barrett, unpublished data; see https://www.affective-science.org/face-set.shtml for further information about the IASLab facial stimulus set). Actors that had the highest average emotion category accuracy ratings and no facial hair were selected and randomly assigned to the different images. The stimuli included open and closed mouth images of anger, calm, disgust, excitement, fear, happiness, neutral, sadness, and surprise for a total of 18 images in each sort. One sort consisted of all 18 images within one individual, while another sort consisted of 18 different individuals (half male and half female, with a male and female for each emotion).

Norming stimuli. 50 undergraduates completed 7-point Likert ratings of valence and arousal for each image (Warriner, Kuperman, & Brysbaert, 2013) and the Evaluative Space Grid for ratings of positivity and negativity (Larsen, et al., 2009).

Design & Procedure

Design & Apparatus. Images were presented on a Dell 24: P2418HT touchscreen monitor using Psychopy (version v1.83.04; Pierce et al., 2019). At the outset of each sorting phase, participants saw all the images to be sorted. The images then disappeared, and the images were presented one at a time in the center of the screen. Participants were instructed to arrange the images that go together or are the same kind of thing by touching the images they wished to move and dragging them to different locations in the grid. In order to ensure that images were clearly visible to participants, images would expand in size (from 315x315 to 140x140) while participants re-arranged the image and return to their original size once placed in the grid. Participants could continue to move each image as many times as they wanted throughout the task. Task design and instructions were based on prior experiments using the spatial arrangement method with children (Unger, et al., 2016).

Demo/ Instruction Phase. In order to introduce participants with the task of sorting images based on similarity, they were exposed to a demo phase where they saw 4 images (soccer ball, basketball, rabbit, and chair) and practiced moving them around on the screen. This allowed a chance to help children better understand the task, as no feedback was given in any other phase of the task.

Practice Phase. Next, participants completed a practice phase in which they were asked to arrange 5 images belonging to different superordinate categories (vehicles: car, bus; animals: squirrel, bird; furniture: table). This practice phase was designed to ensure that participants (especially in the youngest age group) understood the task instructions, by verifying that participants sorted images from other domains of knowledge based on semantic similarity.

Face Sorting Phase 1. Participants sorted 18 facial cues of emotion for actor # 7 in the IASLab Facial Stimuli Set. Our rationale was that sorting only one actor would help participants focus on changes in expression. Children were instructed to think about how the person might be feeling and to use that to decide if the pictures should "go together".

Face Sorting Phase 2. Participants sorted 18 facial cues of emotion for 18 different actors in the IASLab Facial Stimuli Set. By sorting again with multiple actors, this allowed us to examine if similar sorting patterns emerge when there are a variety of different perceptual features that are changing (including expression).

Results

Analytic Approach

Analyses were conducted in R (version 3.6.1; R Development Core Team, 2019). Linear mixed-effects models were fit using the lme4 package (Bates & Maechler, 2009; version 1.1-21). Following the recommendations of Judd et al. (2012), F-values and *p*-values for linear mixed-effects models were obtained using Kenward-Roger approximation of the degrees of freedom. Sorting distances between images were normalized for each participant by scaling distances based on the maximum distance for each participant. See Figure 1 for a visual representation of children's and adult's overall average grouping of faces.

Practice Phase

To assess whether participants in each age group demonstrated understanding of the grid task during the practice phase, we investigated the degree to which participants consistently arranged images belonging to the same superordinate categories (car and bus; squirrel and bird) closer together in space. We computed the average distance between images belonging to the same superordinate category for each participant, and then compared the average distances for item pairs sharing the same category to item pairs from different categories. 3-4-year-olds did not consistently arrange items belonging to the same superordinate category closer together in space, (paired *t*-test: $t(14) = -0.32$, $p = .75$). This suggests that children in this age group were not consistently sorting images according to similarity and may have struggled with the task instructions. All other age groups consistently sorted images belonging to the same category closer together in the grid space (4-5-year-olds: $t(34) = 5.26$, $p < .001$; 5-6-year-olds: $t(25) = 5.85$, $p < .001$; 6-7-year-olds: $t(13) = 5.41$, $p < .001$; adults: $t(39) = 29.10$, $p < .001$).

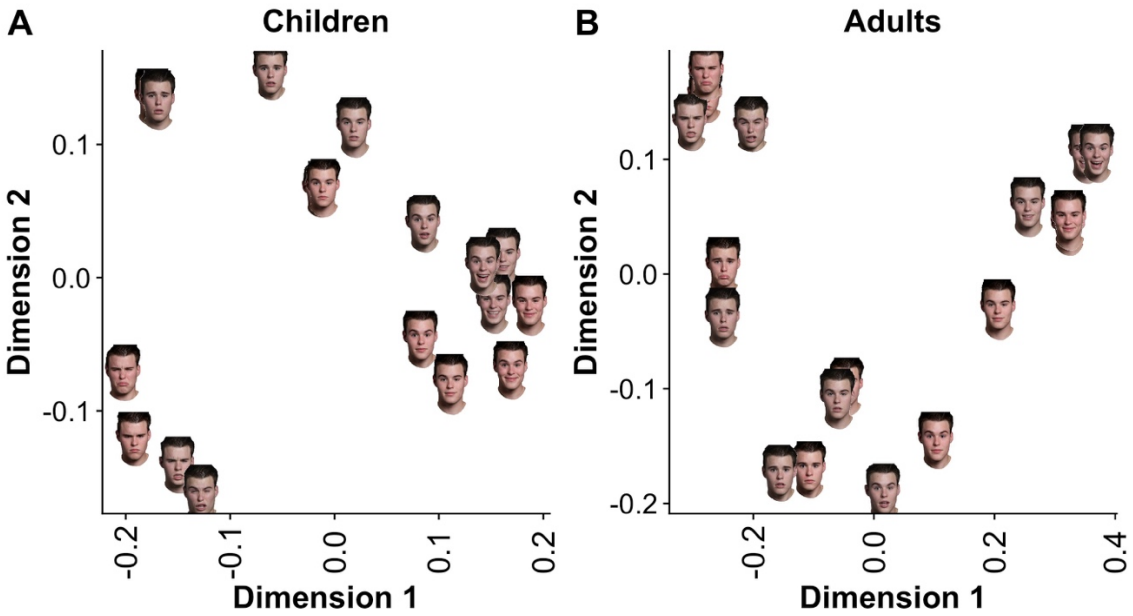

Figure 1. Classical multidimensional scaling solution (2 dimensions) for average sorting distances across (A) all children and (B) all adults in Face Sorting Phase 1 (same actor for each facial cue).

Emotion Categories

To investigate whether participants used emotion categories (e.g., sad, happy, etc.) to structure their placement of facial cues, we computed the average distance between images that shared the same category label versus images that had differing category labels for each participant (see also Unger et al., 2016 for a similar approach). We next fit a series of models to investigate whether participants grouped facial expressions thought to cue the same emotion categories closer together in space than facial expressions thought to cue different emotion categories, and how participants' categorical grouping behavior changed across age (Figure 2).

The use of emotion category information emerges across age in children. We fit a linear mixed-effects model estimating the average distance between item pairs for children from age (in years, as a continuous predictor; mean-centered), the category match for an image pair (same category pair vs. different category pair; centered), and their interaction. We included a by-participant random intercept and a by-participant random slope for category match. The distances between images belonging to the same vs. different emotion categories increased with age, b = -.03, Wald 95% CI = [-.05, -.02], $F(1,86.7) = 19.28$, $p < .001$. As children grow older, they become more likely to sort images belonging to the same emotion category closer together. In follow-up analyses investigating children's sorting behavior in each of our age groups, we found that neither 3-4 year olds (p = .57) nor 4-5 year olds (p = .33) reliably sorted same emotion category images closer together, while, the 5-6 year olds (b = -.07, Wald 95% CI = [-.10, -.04], $F(1,25) = 22.61$, $p < .001$), the 6-7 year olds (b = -.12, Wald 95% CI = [-.15, -.08], $F(1,13) = 45.19$, $p < .001$), and the adults did (b = -.20, Wald 95% CI = [-.21, -.18], $F(1,40) = 487.64$, $p < .001$). Note that while we analyzed results collapsing across sorting phase, there was no evidence that results differed between sorting phases (i.e., no interaction with sorting phase).

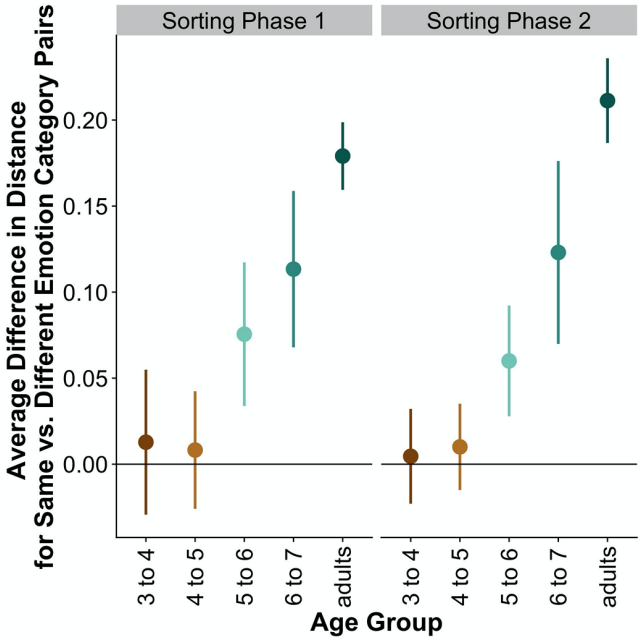

Figure 2. Difference in average distance for items belonging to the same vs. different emotion categories by age. An average value of zero represents no distinction by emotion category. Error bars represent 95% confidence intervals.

What dimensions best predict similarity grouping?

Next, we investigated which dimensions of the emotion expression stimuli best predicted which items participants grouped closer together in space. For each age group we computed the average distance between all stimulus pairs (n= 316 unique pairs) and predicted these distances from different dimensions of the image pairs (valence, arousal, shared emotion category, positivity/negativity). Our analytic approach was to estimate general linear models in which we regressed the average distance between stimulus pairs on stimulus pairs' relatedness on each dimension (e.g., the difference in valence between two images), and to describe how much each dimension aided in explaining variance (by computing ΔR^2). First, we estimated grouping distance from distance in valence, distance in arousal, and whether image pairs shared the same category (0 = different category; 1 = same category). Next, we used a similar approach, but instead of estimating valence as a single dimension ranging from positive to negative, we used the evaluative space grid ratings to estimate positivity and negativity as two separate dimensions.

Table 1: Predicting Sorting Distance from Valence, Arousal, and Shared Emotion Category

Predictor	Estimate	*t*-value	*p*	ΔR^2	Overall R^2
3-4-year-olds					.02
Valence	.004	1.72	.09	.01	
Arousal*	-.009	-2.10	.04	.01	
Same Emotion Category	-.007	-.54	.59	.00	
4-5-year-olds					.13
Valence***	.01	5.73	<.001	.09	
Arousal***	-.02	-4.76	<.001	.07	
Same Emotion Category	-.002	-.16	.87	.00	
5-6-year-olds					.30
Valence***	.03	10.42	<.001	.25	
Arousal**	-.02	-3.19	.002	.02	
Same Emotion Category*	-.04	-2.49	.01	.01	
6-7-year-olds					.38
Valence***	.05	12.25	<.001	.31	
Arousal***	-.02	-3.56	<.001	.03	
Same Emotion Category***	-.07	-3.38	<.001	.02	
Adults					.78
Valence***	.09	30.04	<.001	.65	
Arousal***	-.02	-3.83	<.001	.01	
Same Emotion Category***	-.08	-5.26	<.001	.02	

Note. Asterisks denote significance level, *p < .05; **p < .01; ***p < .001.

Valence, arousal, and shared emotion category. Table 1 summarizes the results of regressing average distance of each image pair on their distance in valence, distance in arousal, and whether images shared the same emotion category. In general, the amount of total variance explained by the model increased steadily across ages, accounting for a significant amount of the error variance in all but the youngest age group (3-4-year-olds: $F(3, 302) = 2.12$, $p = .10$; for all other age groups: $F(3, 302) > 15$, $p < .001$). Distance in valence emerged as (by far) the strongest predictor (in terms of ΔR^2, i.e., the decrement in variance explained by the model when this predictor was omitted) of how closely children grouped two images, emerging as a robust predictor in the 4-5-year-old age group ($t(302) = 5.73$, ΔR^2=.09, $p < .001$) and steadily increasing in ΔR^2 across age groups (Table 1; Figure 3). Arousal was a significant predictor across all age groups (even among 3-4-year-olds), but declined in unique variance explained as children grew older (Figure 3). Consistent with the results from the previous section, shared emotion category emerged as a predictor beginning only with the 5-6-year-old age group, though it explained only a modest amount of unique variance relative to the valence dimension even among adults (ΔR^2=.02 or less among all age groups).

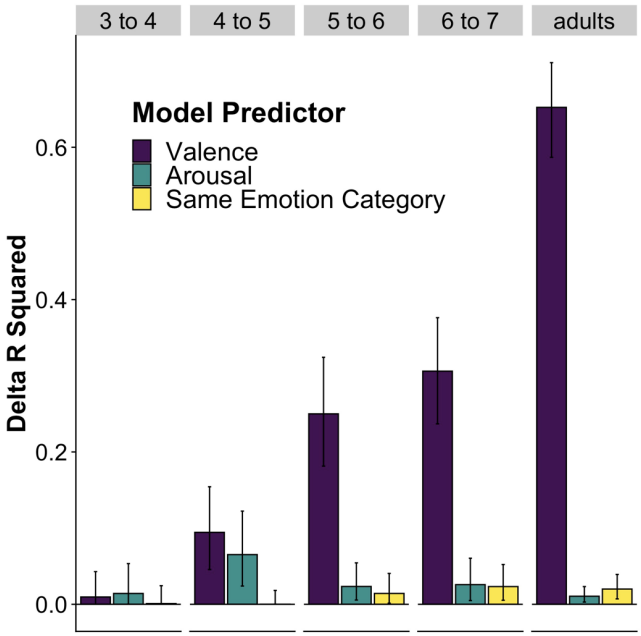

Figure 3. Delta R-squared for each predictor. Error bars represent bootstrapped 95% confidence intervals.

Positivity, negativity, arousal, and shared emotion category. Next, we estimated the influence of different stimulus dimensions on participants' sorting behavior using a similar approach, with the only difference that we replaced distance in valence with distance of positivity and distance of negativity ratings as separate predictors in the regression models for each age group. Note that, as expected, ratings of distance in positivity (r = .92) and negativity (r = .87) were highly correlated with distance in valence ratings. Note that the high correlation between positivity/negativity and valence also precluded us from estimating a model including all 5 predictors due to multicollinearity (and the associated inflation in variance in linear models). Table 2 summarizes the results from the general linear models for each age group. Note that consistent with the importance of valence in explaining sorting behavior in the previous results, the dimensions of positivity and negativity explained substantially more unique variance than the other predictors of distance in arousal and shared emotion category.

We also investigated whether splitting valence into separate dimensions of positivity and negativity led to better

model performance, by comparing the models including valence to the models including positivity and negativity (in addition to arousal and shared emotion category, both included in all models) in each age group. In general, replacing the valence predictor with separate predictors of positivity and negativity led to improved model performance in all but the youngest age group (3-4-year-olds: $F(1, 301) = 3.44$, $p = .06$; 4-5-year-olds: $F(1, 301) = 19.46$, $p < .001$; 5-6-year-olds: $F(1, 301) = 70.66$, $p < .001$; 6-7-year-olds: $F(1, 301) = 78.31$, $p < .001$; adults: $F(1, 301) = 21.88$, $p < .001$), though the gains in model R^2 were most substantial among 5-6- and 6-7-year-olds (compare overall R^2 in Table 1 vs. Table 2).

Table 2: Predicting Sorting Distance from Positivity, Negativity, Arousal, and Shared Emotion Category

Predictor	Estimate	*t*-value	*p*	ΔR^2	Overall R^2
3-4-year-olds					.03
Positivity	-.006	-1.18	.24	.00	
Negativity*	.01	2.42	.02	.02	
Arousal	-.004	-0.76	.45	.00	
Same Emotion Category	-.004	-.30	.76	.00	
4-5-year-olds					.18
Positivity	-.003	-.70	.49	.00	
Negativity***	.02	5.66	<.001	.09	
Arousal*	-.009	-2.43	.02	.02	
Same Emotion Category	.004	.42	.68	.00	
5-6-year-olds					.44
Positivity*	-.01	-2.36	.02	.01	
Negativity***	.07	11.60	<.001	.25	
Arousal	.005	1.04	.30	.00	
Same Emotion Category	-.02	-1.62	.11	.00	
6-7-year-olds					.51
Positivity*	-.02	-2.37	.02	.01	
Negativity***	.10	13.06	<.001	.28	
Arousal	.01	1.17	.24	.00	
Same Emotion Category*	-.05	-2.59	.01	.01	
Adults					.80
Positivity***	.05	8.36	<.001	.05	
Negativity***	.11	15.39	<.001	.16	
Arousal	-.003	-0.54	.59	.00	
Same Emotion Category***	-.08	-4.92	<.001	.02	

Note. Asterisks denote significance level, *$p < .05$; **$p < .01$; ***$p < .001$.

Discussion

We found that children's use of valence, arousal, and emotion labels to categorize facial expressions emerges gradually across development. Valence had the largest impact on adults and children's sorting behaviors, and explained even more variance in behavior when allowed to vary along two dimensions (positivity and negativity). Consistent use of sorting by emotion category emerged around 5-6 years of age, but only explained a small proportion of participant's behaviors.

Crucially, our data suggests that children were not simply misunderstanding the task or re-arranging the items at random. With the exception of the youngest age group (3- to 4-year-olds), children systematically organized images belonging to other domains during the practice phase and sorted images according to broader dimensions such as valence and arousal. We suspect that the failure of the youngest group to systematically sort images in the practice and emotion tasks may be tied to individual differences in language abilities, as children likely needed a basic understanding of "sameness" and "difference" to complete the task. Follow-up analyses investigating individual differences in this youngest age group could also be informative, as some three-year-olds were able to successfully complete the practice task, some appeared to focus on other perceptual features of the categories, such as color, to guide their decisions, and some appeared to sort at random.

Use of Emotion Categories Emerges Gradually

Adult participants showed clear evidence of using emotion categories to guide their sorting decisions: adults tended to sort images belonging to the same emotion category (e.g., two face images that are typically classified as cuing the emotion "sadness") closer together in space than two images that belonged to different emotion categories. By contrast, this tendency emerged only gradually in children's sorting behavior. We found little evidence that children were consistently using emotion categories in grouping images until around the age of 5-6 years of age, and the sorting behavior in our oldest age group (6-7 years of age) still showed clear differences to the category-based sorting behavior observed in adults.

What Dimensions Guide Similarity Judgments?

Our finding that valence is a large part of sorting behavior replicates prior findings tied to individual's representations of emotion words (Jackson et al., 2019; Nook, Sasse, Lambert, McLaughlin, & Somerville, 2017). While arousal contributed to adult's and children's sorting of facial categories, it explained a much smaller portion of variance than valence. Arousal may be a more important component when one has more information about an individual's facial expressivity. Personality, gender, and culture can influence facial expressivity and both adults and children use this information to update facial emotion categories (Plate, Wood, Woodard, & Pollak, 2018).

The variance explained by arousal disappearing when valence is allowed to vary in two dimensions (positivity and negativity) is also notable. This suggests that arousal may be indexing some of the covariation in positivity and negativity, rather than capturing unique variance tied to activation.

Limitations and Future Directions

Individual differences. In the reported analyses, we thus far consider only group-wide differences based on age. Thus, some of the differences across groups may be due in part to greater variation among individuals in some age groups compared to others. In future analyses, we aim to investigate individual differences in children's sorting decisions and how they relate to individual differences among other dimensions, including demographic factors and children's emotion

vocabulary knowledge (collected in conjunction with the current task).

Generalizing across stimuli and participants. The current work represents an exploratory first step in implementing a novel method of measuring the development of children's judgements of facial expressions of emotion. In future work, we aim to expand the current method beyond the small number of facial stimuli tested in the current experiments and to larger groups of participants and age ranges.

Distinguishing developmental changes in task fluency from emotion development. A central challenge for future work will be developing analytic methods for distinguishing changes in children's task-specific strategies from changes in the underlying representations children use when interpreting facial expressions of emotion. The current approach reported here partially addresses this difficulty by (a) validating that children interpret the general goal of the task similarly across ages by comparing sorting behavior in other domains during the practice phase and (b) by estimating changes in the extent to which participants rely on specific stimulus dimensions within each age group. Future work will seek to further address this problem by investigating how children and adults approach the task across different semantic domains, and by studying the effect of experience-based interventions targeted at changing children's perception and knowledge of facial expressions of emotion (Unger & Fisher, 2019).

Conclusions

The spatial arrangement task is a valuable method that can help us better understand children's emotional development and representations of emotion knowledge. We found that the task showed meaningful change in children's use of valence, arousal, and emotion labels in categorizing facial expressions of emotion. Crucially, children appear to slowly and continually learn to detect the dimensions of valence and arousal in others' facial cues. These improvements likely precede improvements in the use of emotion labels, and future directions could further explore how development in these two areas may be related.

References

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes.* Version 1.1-21.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

Gendron, M., Lindquist, K. A., & Barrett, L. F. (2006). The IASLab Face Set. Unpublished data. https://www.affective-science.org/face-set.shtml

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381-386.

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, *28*(2), 219-236.

Jackson, J. C., Watts, J., Henry, T. R., List, J. M., Forkel, R., Mucha, P. J., ... & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517-1522.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, *103*(1), 54-69.

Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkley, L. C., & Cacioppo, J. T. (2009). The evaluative space grid: A single-item measure of positivity and negativity. *Cognition and Emotion*, *23*(3), 453-480.

Montirosso, R., Peverelli, M., Frigerio, E., Crespi, M., & Borgatti, R. (2010). The development of dynamic facial expression recognition at different intensities in 4-to 18-year-olds. *Social Development*, *19*(1), 71-92.

Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2017). Increasing verbal knowledge mediates development of multidimensional emotion representations. *Nature human behaviour*, *1*(12), 881-889.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods.* 10.3758/s13428-018-01193-y

Plate, R. C., Wood, A., Woodard, K., & Pollak, S. D. (2018). Probabilistic learning of emotion categories. *Journal of Experimental Psychology: General*.

R Development Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.

Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., & MacLellan, C. J. (2016). Developmental changes in semantic knowledge organization. *Journal of Experimental Child Psychology*, *146*, 202-222.

Unger, L., & Fisher, A. V. (2019). Rapid, experience-related changes in the organization of children's semantic knowledge. *Journal of experimental child psychology*, *179*, 1-22.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191-1207.

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, *23*(2), 291-312.