Children's Representation of Facial Cues of Emotion Across Development

By

Kristina Woodard

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 12/10/2021

The dissertation is approved by the following members of the Final Oral Committee:

Paula M. Niedenthal, Professor, Psychology
Martha W. Alibali, Professor, Psychology
Gary Lupyan, Professor, Psychology
Haley A. Vlach, Associate Professor, Educational Psychology

Table of Contents

Acknowledgements……..……..iii

Abstract……...…iv

Introduction…..…….……1

Variability in Facial Cues of Emotion…….……1

Representing Children's Knowledge of Facial Cues of Emotion…….…….…..……4

Language and the Development of Emotion Concepts……7

The Spatial Arrangement Method…….……..…….……9

Categories and Dimensions Underlying Emotion Understanding…….……10

Method…..…….....…12

Participants…….……12

Stimuli……14

Design and Procedure……15

Results…..……..…….……18

Practice Phase ……18

Comparing Dimensions of Affect and Categories in Sorting Behaviors …….…….20

Emotion Category……..…….…….…….…….……20

Dimensions of Affect…….…….…..…….…….…..……24

Comparing Dimensions of Affect and Emotion Categories……..……26

Bottom-up Assessment of Facial Cues of Emotion……...…..…….…….……29

Multidimensional Scaling……..…….…….…….…….……29

Hierarchical Clustering……..…….…….…….…….……33

Verbal Fluency Task……..…….…….…….…….……..….39

Descriptive Analysis……..…….…….…….……...……40

Comparison to CHILDES……..…….…….…….…….……41

Analysis of Verbal Fluency and SpAM…….……...…….……44

Multidimensional Scaling……..…….…….…….……47

Hierarchical Clustering……..…….…….…….…….……48

General Discussion…..……..…..50

Nuances in the Use of Valence……..…….…...……50

The Limits of Arousal……..……..……51

Verbal Fluency Task……..……...……53

What Changes in the Representation of Facial Cues of Emotion? ……..……..…54

Task Understanding……..…….……55

The Role of Dimensions of Affect……..…….……..…55

The Role of Emotion Categories……..…….……..…56

Children's Conceptual Development……..…….……..…57

Perceptual Learning from Distributions……..…….……..…58

Implications for Theories of Emotion……..…….……59

Limitations…..…….……..….60

Future Directions…..……...…62

Conclusions…..……...…63

References…..…….…..…..65

Appendix A on Stimulus Ratings…..…….…..…..83

Appendix B on Gender…..…….……84

Acknowledgements

I would like to acknowledge the many who contributed to this dissertation. Thank you to the members of my committee, Paula Niedenthal, Martha Alibali, Gary Lupyan, and Haley Vlach for their feedback and the larger roles they have played in my academic career. I would also like to thank the many research assistants who ran participants and motivated me with their fresh ideas and insights, particularly Sarah Fieweger, Stacey Sukoff, Quentin Wedderburn, Chloe Stevens, Alex Henoch, and Julian Bok, and the children and families who contributed their valuable time to participate.

Thank you also to my incredible academic and emotional support system, particularly Martin Zettersten, Rista Plate, Karen Smith, Sarah Brown, Desia Bacon, Alyssa Lovely, Lillian Xu, Lucia Pozzan, and my former undergraduate advisor John Trueswell. My husband Spencer, cat (Mochi!!), and family have been my well of support throughout this process.

Abstract

The present study examined how children spontaneously represent facial cues associated with emotion. 107 three- to six-year-old children (gender: 48 M, 59 F; race/ethnicity: 9.3% Asian, 84.1% White, 6.5% multiple) and 40 adults (gender: 10 M, 30 F; mean age = 18.8; race/ethnicity: 10% Hispanic, 30% Asian, 2.5% Black, 57.5% White) sorted emotion cues in a spatial arrangement method that assesses emotion knowledge without reliance on emotion vocabulary. Using supervised and unsupervised analyses, the study found evidence both for continuities and gradual changes in children's emotion knowledge compared to adults. Knowledge of emotions develops through an incremental learning process in which children change their representations of emotion using combinations of factors—particularly valence—that are weighted differently across development.

Keywords: emotion knowledge; categorization; face processing; free sorting; development

Running Head: Children's understanding of facial cues

Children's Representation of Facial Cues of Emotion Across Development

Historically, many Western philosophers discussed emotions as separate from cognition, irrational, animal-like, and in need of control by reason. Yet, those with deficits in emotion recognition and understanding struggle to form social relationships and often suffer from other psychopathologies. This suggests that emotion is not irrational, but a part of cognition, and a skill essential for social bonds and wellbeing (Dukes et al., 2021; Ochsner & Phelps, 2007; Pessoa, 2008).

For children, emotions are a rich source of information that can be utilized to formulate predictions about what is likely to occur in their environments. Faces are a particularly salient early cue for social and emotional learning and form a large part of children's early visual experience (e.g., Fausey, et al., 2016; Jayaraman, et al., 2017; Jayaraman & Smith, 2019). For instance, facial movements from others, in combination with other contextual information, help children understand whether their actions are approved of by their social partners or caregivers, whether they should approach or avoid persons, and whether an environment is safe (Walle et al., 2017).

Variability in Facial Cues of Emotion

While knowledge of facial cues of emotion is sometimes thought to be "innate," children face a difficult learning problem. The human face contains over 40 muscles capable of producing over 16,000 different muscular combinations (Srinivasan & Martinez, 2018), and provides information on a variety of topics including identity, age, race, gender, and speech perception. Furthermore, children encounter substantial variability in the facial cues of emotion that they experience in their environments (for a detailed review see Barrett et al., 2019). Different people

might convey similar emotions with different facial movements, or with varying levels of subtlety, intensity, or degree of muscular movement (Cordaro et al., 2018; Kring & Gordon, 1998). And the same person might convey similar feelings differently at different points in time or in different contexts. Yet across this variability, children develop concepts to systematically distinguish between emotional states.

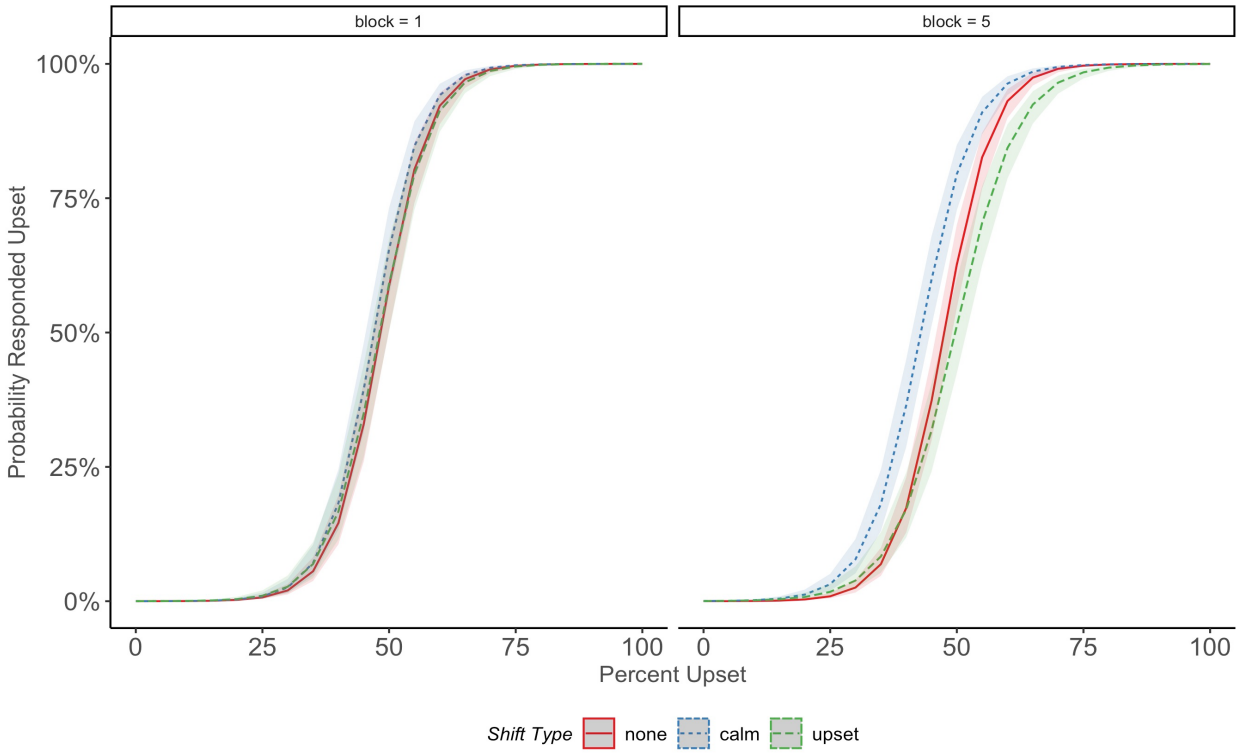
My research has focused on understanding how children learn to navigate and use information from individual differences in facial cues in order to more accurately predict how others might feel and react. One interpretation of the variability encountered in facial cues is that the extensive variability makes this perceptual information unhelpful. Children must instead rely on language in order to learn about emotion categories (e.g., Hoemann et al., 2019). An alternative possibility is that the natural variability in facial cues is itself an important source of learning. This variability is used to learn about emotion categories and adjust to individual differences in expressivity. In this instance, distributional properties of the perceptual input are used to adapt to different speakers and contexts (Kleinschmidt & Jaeger, 2015).

Much of my work on adjusting to individual differences in expressivity supports this latter possibility. For instance, children and adults rapidly adjust their categorization of facial (Plate, Wood, et al., 2019; Woodard, Plate, & Pollak, 2021) and vocal cues (Woodard, Plate, Morningstar, et al., 2021) based upon distributional properties of the perceptual input encountered. When a speaker's facial or vocal intensity is limited such that they never express "maximal" negative arousal, cues that were previously categorized as calm are categorized as upset. In other words, when exposed to less expressive speakers, people have lower thresholds for detecting emotion in the face and voice. Likewise, if a speaker's expressive range is more intense, cues that were previously categorized as upset are categorized as calm. Perceivers adapt

to highly expressive speakers by increasing their threshold for detecting emotion in the face and voice (see Figure 1). These adjustments occur across a range of facial and vocal stimuli (both verbal and nonverbal) and for multiple speakers and negatively-valenced emotions. Thus, the variability of emotional cues appears to be an important source of learning and could help to explain the formation and maintenance of emotion "dialects" and differences in cross-cultural emotion recognition (Laukka & Elfenbein, 2021).

Figure 1

Children Adjust to the Facial Expressivity of Different Actors



Note. Children adjust their boundary for labeling different actors "upset" based on the distributional information encountered. Overtime, anger is identified earlier in less expressive actors ('calm' condition), and later in more expressive actors ('upset' condition).

Representing Children's Knowledge of Facial Cues of Emotion

While the above work highlights how variability in perceptual cues of emotion drives learning, it fails to give a broader picture of how different perceptual cues may be related, and how underlying categories and dimensions might drive this learning. Emotional cues are interrelated and have different degrees of similarity. For instance, facial cues of anger and disgust are often viewed as more similar to one another (Widen & Russell, 2010a) than cues of anger and sadness. These relationships suggest that much of emotion experience and expression may be best represented by a semantic network (see Cowen & Keltner, 2021 for a discussion on semantic spaces of emotion experience and expression).

A semantic network is a representation of the relationships between different concepts. That is, rather than knowledge being isolated information (e.g., cats have ears and whiskers, and are pets), conceptual knowledge also includes an understanding of relationships between different concepts (e.g., cats are pets like dogs, both cats and seals have whiskers). The study of children's semantic development in non-emotion domains provides information on how—despite the vastness and variety of things encountered—meaningful clusters and dimensions are learned and derived from experience (e.g., distinguishing plants versus animals versus buildings). Experience in particular plays an important role in semantic differentiation (for a discussion of language as experience see the next section: *Language and the Development of Emotion Concepts*). Children show differences in the structure of their semantic knowledge based on experiences with plants and animals (Unger & Fisher 2019; Vales, States, et al., 2020), and individual expertise (Gobbo & Chi, 1986; Winkler-Rhoades et al., 2010).

In a similar way, emotion concepts may be shaped by experience and change over time. However, much of the work on children's emotion knowledge is not able to inform how facial

expression understanding develops over time. The primary challenge to understanding this development concerns the difficulty in accurately assessing what children are perceiving when they are exposed to stimuli such as facial configurations (Barrett et al., 2019). Much of the data used to understand the structure of young children's emotion knowledge relies upon children's production and comprehension of emotion labels (see Ruba & Pollak, 2020, in particular the sections on verbal-response paradigms). The most commonly used approaches in this field involve asking children to generate a verbal label to describe a facial stimulus such as "*What is this person feeling?*" (Nelson & Russell, 2011; Widen & Russell, 2003). Other common methods include sorting images into labeled piles (Hoemann, et al., 2021; Matthews, et al., 2020), confirming whether labels match an image displayed with prompts such as "*Is this person feeling sad?"* (Widen & Russell, 2008), or selecting a stimulus from an array of predetermined response options (Chronaki et al., 2015; Leitzke & Pollak, 2016; Pollak & Sinha, 2002). In the latter case, children are asked to either select a label to match a face (*Is this face angry, happy, or scared?*) or pick a face to match a label (*Choose the face that looks happy*).

These approaches share three key limitations. First, these methods are constrained by the emotion categories determined by the researcher: the researcher selects stimuli they believe represent "happy" or "sad" and accept only happy and sad as correct answers for those stimulus items. This approach can reveal the degree to which children successfully align their responses with the (adult) researcher's view of emotion (e.g., labeling a "sad" face as "sad" given the options "happy", "angry", and "sad"), but provide limited insight about a child's own construal of the faces, which might not map onto any of the labels or categories that the researcher selected.

Second, verbal-response methods equate knowledge of an emotion vocabulary word with a child's use of perceptual information. This assumption can underestimate what children actually know about emotion. Many emotion words are not learned until later in development (Baron-Cohen et al., 2010), word comprehension often precedes word production (Bergelson & Swingley, 2012, 2013), and social referencing paradigms indicate that infants are adaptively using facial movements to guide their behavior long before expressive emotion vocabulary is present (Walden & Ogan, 1988). For these reasons, it is unsound to assume that a child who cannot produce, comprehend, or use a word such as "scared" does not know something about the concept of fear or threat. Furthermore, seemingly simple emotion words change in abstraction across development (Nook et al., 2020) – as is the case for objects (Ameel, et al., 2008) and other abstract concepts like time (Tillman & Barner, 2015)—making it difficult to interpret whether children and adults even mean the same thing when using a labels such as "mad", let alone complex ideas such as love or shame.

Third, most extant procedures were not designed to provide information about how children think about the relations among emotion cues. Past work has explored the dimensional and categorical mappings of emotion in adults (Cowen & Keltner, 2017); however, it is still unclear what these relations might look like in children, and how they develop. Some kinds of relations can be inferred through patterns of errors observed in verbal-response paradigms—such as the consistency of children's confusion about anger versus disgust (Leitzke & Pollak, 2016; Widen & Russell, 2010a). Yet, for the most part, information about how children perceive and think about underlying relations among emotion cues is limited. This limitation also reflects a broader problem in emotion research: interpretations of children's "errors" are often predicated on the assumption that deviations from the researcher's pre-determined label for an emotion

stimulus are incorrect—that is, if the researcher has labelled a stimulus face as “sad”, other interpretations or reactions to those stimuli are coded as errors.

These limitations have made it hard to uncover the manner in which children’s understanding of facial cues of emotion unfolds. One possibility is that children’s knowledge of facial cues begins more broadly (e.g., two categories of positivity and negativity that are refined over time). Another possibility is that children’s knowledge of facial cues begins more narrowly (discrete emotion knowledge like sad/angry/happy) and broadens over time (i.e., grouped by superordinate categories like positivity or negativity, see Ruba & Repacholi, 2020). Studies with infants support the narrow to broad pattern (e.g., discrete emotions followed by valence categories; White et al., 2019), with the caveat that labels may help infants to learn broader patterns (Ruba, et al., 2020). Conversely, studies with young children support the broad to narrow pattern. Children first use broad, primarily valence-based distinctions, and with greater experience, draw more fine-grained distinctions that use emotion category information (e.g., Matthews et al., 2020; Widen, 2013; Widen & Russell, 2010b). The present study aims to contribute to this debate by using a task that does not require children to make binary choices, does not rely heavily on emotion vocabulary, and does not force children to sort into predetermined emotion categories.

Language and the Development of Emotion Concepts

While the previous section highlighted limitations in how emotion vocabulary has been utilized in studies of children’s emotion understanding, the section is not meant to imply that language plays no role in the learning of facial cues and emotion concepts. Language is still an important piece of children’s emotion experience and development, and increasingly believed to play an important role in guiding and forming conceptual knowledge about emotion (e.g.,

Hoemann, et al., 2019; Lindquist, 2021; Shablack & Lindquist, 2019). For instance, parents' use of emotion language predicts children's emotion understanding (LaBounty, et al., 2008), emotion language helps children with emotion regulation (Cole, et al., 2010), children with a more advanced emotion vocabulary make fewer errors on emotion sorting tasks (Matthews et al., 2020), and labels help children to form superordinate valence categories about faces that they might not otherwise (Ruba, et al., 2020). This relationship extends to adulthood – in which adults with more fiction reading experience (a proxy for experience with complex emotion terms) had better emotion recognition (Schwering et al., 2021).

In many ways – the current discourse on the role of language in children's emotion understanding is an expansion of the debate on the role of language in children's word learning generally and in adults' "language-augmented cognition" (Lupyan, 2016). Part of this discourse involves whether—for young children—words serve a supervisory role as "invitations to form categories" (Waxman & Markow, 1995), or whether words initially serve an unsupervised role as another perceptual feature of a concept with no privileged status (Sloutsky & Fisher, 2004). In this view, language's impact on categorization depends on a label's salience and redundancy with other features.

Numerous infant studies support language's supervisory role in early category learning. For example, if two objects are called "dax", 12-month-old children treat both objects as belonging to the same category. However, if one object is called "dax" and another "wug", 12-month-olds treat them as belonging to two different categories (Waxman & Braun, 2005). However, category learning is not a uniquely human process – animals like pigeons can be trained to learn categories too (see Zentall, et al., 2008 for a review). Therefore, an alternative hypothesis is that many of language's facilitating effects occur because auditory information

serves as a salient perceptual cue. For instance, when words do not correlate with other perceptual information (e.g., words are randomly rather than systematically assigned), children learn no categories (Plunkett, et al., 2008), suggesting that labels are helpful only if they correlate with other features (Plunkett, 2011). The state of this debate (and the above work on children's use of variability in categorizing the emotions of actors with different expressivity) suggests that facial cues of emotion still play an important role in children's emotion knowledge, including children's learning of emotion words.

The Spatial Arrangement Method

The present study sought to understand how children represent facial cues of emotion, without introducing verbal labels or assumptions about the accuracy of participants' responses. To do so, I adapted the Spatial Arrangement Method (SpAM) developed by Goldstone (1994a) that has been successfully used and validated alongside more traditional pairwise similarity judgments with both adults (e.g., Hout, Goldinger, et al., 2013; Hout et al., 2016; Hout & Goldinger, 2016) and children (e.g., Unger et al., 2016). In SpAM, participants freely sort images according to the extent to which they perceive stimuli as semantically related. The task was initially designed to be a quick way to collect similarity ratings, as large item sets quickly made pairwise similarity ratings infeasible. Over time, the applications and uses for similarity data have expanded. For instance, similarity plays a role in generalization (Shepard, 1987), categorization (Goldstone, 1994b) and children's category formation (Sloutsky, 2003). Thus, the results of SpAM allow insight into the dimensions and clusters that drive similarity (Koch, et al., 2020) – even for high-dimensional, conceptual categories (Coburn et al., 2019; Richie et al., 2020;). More recently, similarity has been used with children as a proxy for the representation of semantic knowledge itself (Vales, Stevens, et al., 2020) that changes in response to experience

(Unger & Fisher 2019; Vales, States, et al., 2020). In this way, the present study aims to use judgements about the similarity of facial cues to give insight into the underlying dimensions and categories of how children represent facial cues of emotion. Additionally, because SpAM uses graded similarity judgments (i.e., the distance between images) rather than the accuracy of different labels ("correct", "incorrect"), the method should help to better characterize patterns of change across development.

In addition to SpAM, the present study also employed a verbal fluency task involving listing words from an associated category (e.g., "say all of the animal words you know", "say all of the emotion words you know"). Verbal fluency tasks can give insight to structure of semantic knowledge (e.g., Zemla & Austerweil, 2018; Zemla, et al., 2020); however, there are limitations to using this task with children. The same words need to occur frequently enough across multiple participants, which can be more difficult to achieve when working with young children (Unger et al., 2016). Nevertheless, children's verbal fluency of emotion words could be a gross index of children's conceptual knowledge of emotion and allows for comparisons to their sorting of facial cues in SpAM.

Categories and Dimensions Underlying Emotion Understanding

The present study sought to use SpAM and verbal fluency tasks to examine broader patterns in the underlying dimensions and categories children may use to represent facial cues of emotion. The categories and dimensions examined were selected based on the most widely explored hypotheses in children's emotional development. The longest standing theory about the structure of emotion from early infancy was proposed by Katharine Bridges (1932), who observed that children begin by fluctuating between a resting state of calm with punctuated states of distress. This view was the basis of contemporary theories that human understanding of

emotions begins with differentiation between distress/lack of distress, and becomes elaborated over time into fine-grained emotion categories (Nook & Somerville, 2019; Widen, 2013; Widen & Russell, 2008, 2010b). These theories leave unresolved how children organize and represent the range of perceptual features they encounter and how this becomes elaborated over development.

The concepts most frequently used to refer to the initial building blocks of emotion experience and perception are valence and arousal (e.g., Bliss-Moreau et al., 2020; Russell, 2003). Valence (positivity/negativity) can be conceptualized either as bipolar (a single scale from positive to negative with a neutral midpoint) or bivariate (two orthogonal scales of positivity and negativity; Larsen et al., 2009; Mattek et al., 2020). The dimension of arousal captures low to high activity or engagement. Other theories propose that key physical features such as open or closed mouths form not only the basis of face perception, but also emotion reasoning (Caron et al., 1985). And still other views maintain that children have a rudimentary sense of a limited set of emotion categories that they use to understand facial configurations (Izard, 2007; Leppänen & Nelson, 2009). Historical and anthropological perspectives have emphasized language as a key building block of emotion (Harré, 1986; Lutz & White, 1986), a view that has recently re-emerged (Hoemann et al., 2019; Lindquist, 2021; Nook et al., 2020).

The present study tested predictions that follow from extant theories about the emergence of human emotion, including the possibilities that (a) children use emotion categories (Izard, 2007; Keltner et al., 2019), resulting in facial configurations with the same category label being judged more similarly (i.e., placed more closely together) than those with different category labels across development; (b) children use continuous dimensions including bipolar valence and arousal (Russell, 2003), resulting in facial configurations with more similar bipolar valence and

arousal ratings being judged more similarly; (c) children use valence in a bivariate manner (Larsen et al., 2009), resulting in facial configurations with more similar their bivariate valence ratings being judged more similarly; and (d) children use a combination of these aforementioned features, which predicts that the valence (bivariate and bipolar), arousal, and emotion categories will all explain unique variance in how closely children make similarity judgements about facial cues. It is also likely that with learning and maturation, representation of facial cues of emotion changes. To explore this possibility, I tested children as young as age 3;0 (the earliest age I conjectured children may be able to use this method) through age 6;11 (when children label many emotions similarly to adults) and compared children's behaviors to those of adults. I approached the data in two distinct ways: (1) a top-down, supervised approach to test the extent to which predefined emotion categories and dimensions predict sorting behavior in SpAM, and (2) a bottom-up, unsupervised approach examining participants' behavior without prescribing primacy to any given theory or any specific dimension.

Method

Participants

I recruited 107 children (age range 3;0-6;11 years, mean = 5.0, SD = 1.1; 48 M, 59 F; race/ethnicity: 6.5% more than one race, 84.1% White, 9.3% Asian) and 40 adults (age range: 18-21 years, mean age = 18.8, SD = 0.7; 10 M, 30 F; race/ethnicity: 10% Hispanic, 30% Asian, 2.5% Black, 57.5% White) from the community in a large Midwestern city (Madison, Wisconsin). One 4-year-old child completed only the practice phase and the Same Individual Sort, and one additional 4-year-old child completed only the practice phase. I aimed to have 30 children in each age bin but had to terminate data collection early because of the COVID-19

outbreak; the final sample reported here includes 21 3-year-olds, 35 4-year-olds, 28 5-year-olds, and 23 6-year-olds. 20 participants per subgroup has provided sufficient power for most cluster analysis techniques (Dalmaijer et al., 2020), and the sample size is comparable to those in past studies using the spatial arrangement method with children (n = 18 per group, Unger et al., 2016).

A subset of children also participated in the verbal fluency task. The verbal fluency task was quite challenging for children, and many opted not to participate at all (N=13, gender: 7 F, 6 M; age bins: 6 three-year-olds, 4 four-year-olds, 1 five-year-old, 2 six-year-olds; mean age = 4.46 years). Due to recorder malfunction, an additional four children were excluded (gender: 4 M; age bins: 2 three-year-olds, 1 four-year-old, 1 five-year-old; mean age = 4.12 years). Of the remaining 90 participants, 15 (gender: 9 F, 6 M; age bins: 8 three-year-olds, 6 four-year-olds, 1 five-year-old; mean age = 3.88 years) did not complete the entire task, leaving 75 participants (gender: 43 F, 32 M; age bins: 5 three-year-olds, 24 four-year-olds, 25 five-year-olds, 21 six-year-olds; mean age = 5.36 years). Overall, younger children were more likely to not complete the task and to not participate in the task at all.

Stimuli

Stimuli were drawn from the Interdisciplinary Affective Science Laboratory (IASLab) Facial Stimuli Set.[1] I selected actors with the highest average accuracy ratings and no facial hair, and then randomly assigned each actor to a different emotion category. The stimuli were designated by IASLab as open and closed mouth versions of anger, calm, disgust, excitement,

[1] Development of the Interdisciplinary Affective Science Laboratory (IASLab) Face Set was supported by the National Institutes of Health Director's Pioneer Award (DP1OD003312) to Lisa Feldman Barrett. Gendron, M., Lindquist, K. A., & Barrett, L. F. (unpublished data). More information available online at https://www.affective-science.org/face-set.shtml.

fear, happiness, neutral, sadness, and surprise for a total of 18 images in each sorting condition. I selected these expressions because they are the most commonly tested categories in children's emotional development (the basic emotions and neutrality), and because the facial set includes a range of more positive emotions (calm, happy, excited) which would allow the study to better examine the dimensions of valence and arousal. To test for the robustness of any possible effects, each participant completed two sorting conditions. One sorting condition consisted of 18 different facial configurations posed by the same individual; the other sorting condition consisted of 18 different individuals (half male and half female, with a male and female for each emotion). In this manner, the Same Individual condition reveals how participants construe different facial configurations from one individual, whereas the Different Individual condition reflects a generalization across individual actors, allowing examination of whether similar sorting patterns emerge when a variety of different perceptual features are changing (facial cue, identity, race, and gender).

Ratings of stimuli

Fifty undergraduates who did not participate in the sorting task completed ratings of bipolar valence, bivariate valence (i.e., ratings of positivity and negativity), and arousal for each of the 36 images. Ratings of the stimuli were collected for use as predictors in the analyses of sorting behavior. For each image, participants completed 7-point Likert ratings of bipolar valence and arousal (as in Warriner et al., 2013) and the Evaluative Space Grid for bivariate valence (ESG; Larsen et al., 2009). Valence is often treated as a bipolar measure ranging from negative at one pole to positive at the other with a neutral midpoint. However, bivariate valence—representing positivity and negativity in a two-dimensional space—has been found to

more accurately capture emotional experience (Larsen & McGraw, 2011; Watson et al., 1999). Traditional bipolar valence scales also pose interpretive challenges: scores in the middle of the scale could indicate that the individual perceives the stimulus as neither positive nor negative (indifference, neutrality), that the individual perceives a mix of positivity and negativity (ambivalence, multiple emotions), or that the perceiver is uncertain (a stimulus could be either positive or negative depending upon the context). The ESG method disentangles these possibilities by presenting participants with a square depicting a 5-point positivity scale on one axis and a 5-point negativity scale on the other, allowing participants to select where the stimuli fall along both dimensions. Additional details on stimuli ratings are available in Appendix A.

Design and Procedure

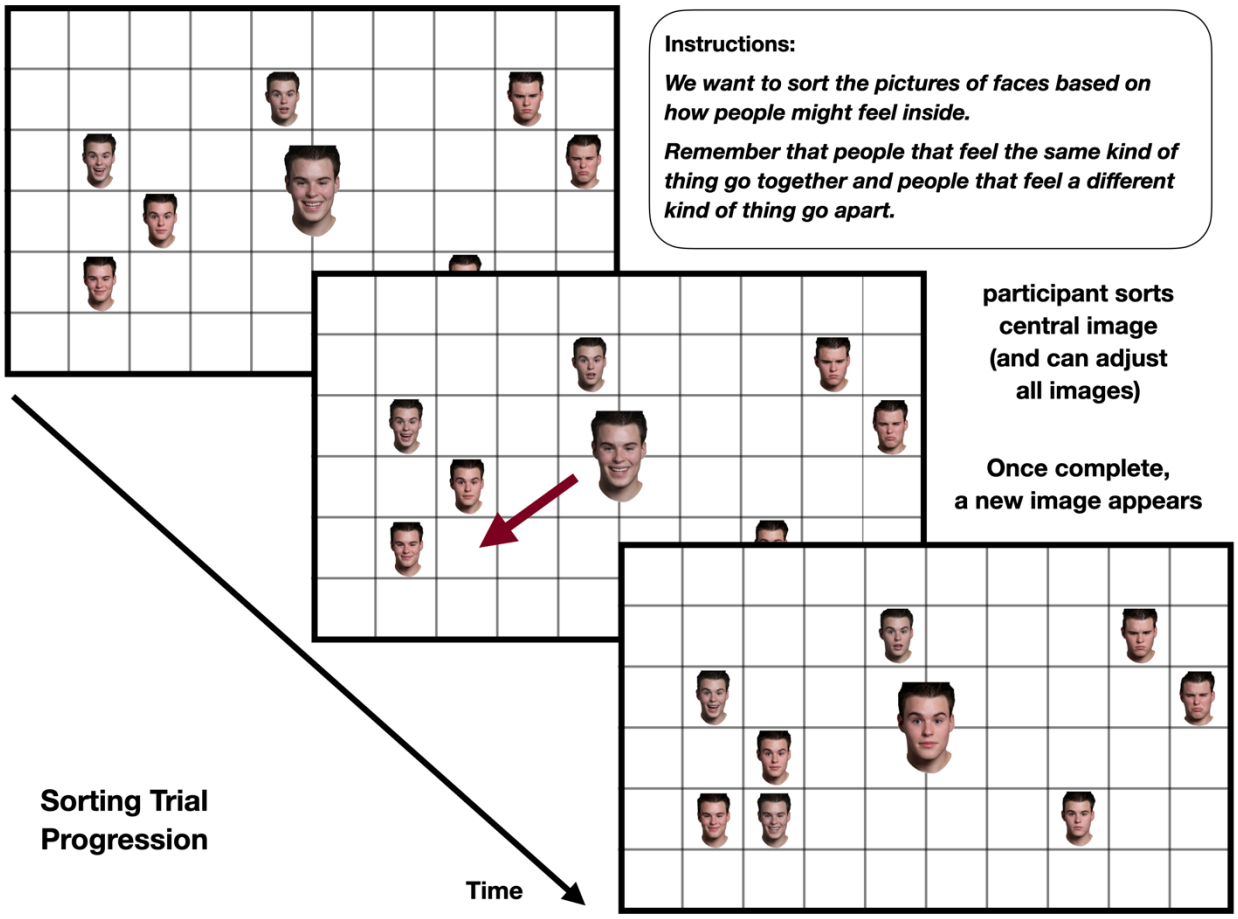
Images were presented on a Dell 24: P2418HT touchscreen monitor using PsychoPy [version v1.83.04; Peirce et al., 2019]. To introduce participants to the task, they saw 4 images (soccer ball, basketball, rabbit, and chair) and practiced moving them around on the screen. The grid had no labels or axes, so participants were not sorting onto a predefined space. The experiment began with a practice phase in which participants arranged 5 images (car, bus, squirrel, bird, table). Children did not receive any feedback during this phase, which allowed for an assessment of how participants approached the task independent of the emotion stimuli (by, for example, grouping the animals together or, as one child explained, grouping the squirrel and the table together because “they both have legs”). For the next two conditions, participants saw faces and were instructed to think about how the person might be feeling, and that people feeling the same kind of thing should go together. Participants then completed a *Same Individual Condition* in which they sorted 18 facial cues of emotion for one actor (Male # 7). Next, participants completed a *Different Individual Condition*, this time sorting 18 facial configurations

posed by 18 different actors (Females: # 1, 4, 7, 10, 13, 14, 15, 17, 22; Males: # 2, 3, 4, 5, 8, 12, 14, 15, 17).

At the outset of each sorting condition, participants saw all of the images to be sorted. The images then disappeared, and each image was presented one at a time in the center of the screen in a randomized order for each participant (Figure 2). Participants were able to arrange the images by touching and dragging them to any location on the grid. For the practice phase, participants were instructed to arrange the images so that "things that are of the same kind of thing go together and things that are different or not the same kind of thing go apart". For the facial sorts, participants were told to "sort the pictures of faces based on how people might feel inside" and that "people that feel the same kind of thing go together and people that feel a different kind of thing go apart". Participants could continue to move each image throughout the task, as all images remained viewable after they appeared. In order to ensure that images were clearly visible to participants, images would expand in size (from 140x140 to 315x315 pixels) while participants touched them to move the image, and then returned to their original size once placed in the grid. Once child participants were no longer moving any images, the experimenter asked if they were ready for the next picture. Adult participants were able to control when the next image would appear themselves by using the spacebar. Adults and children received the same task instructions, though adults were also informed at the beginning that the instructions were designed to also be appropriate for younger participants.

Figure 2

Example of the Structure of the Trials in the Task for the Same Individual Sort



Verbal Fluency Task

After the sorting tasks, children completed a verbal fluency task in which they were instructed to say, "All of the animals that they know" and then "All of the feeling or emotion words that they know". The task ended when children could not think of any more words to say. Because this task was so challenging for participants, the experimenter gave some feedback during the task (nodding, smiling, "mhm", "yeah", "any more?") to help reduce stress.

Results

Analyses were conducted in R (version 4.0.3; R Development Core Team, 2020), using the tidyverse package (Wickham et al., 2019). I fit linear mixed-effects models using the lme4 package (Bates et al., 2015), created dendrograms using ggdendro (De Vries & Ripley, 2020), ggtree (Yu, 2020) and dendextend (Galili, 2015), ran repeated measures correlations using rmcorr (Bakdash & Marusich, 2017, 2021), and examined the CHILDES corpus using the childesr package (Braginsky et al., 2020; Sanchez et al., 2019). Following the recommendations of Luke (2017), *F*-values and *p*-values for linear mixed-effects models were obtained using the Satterthwaite approximation of the degrees of freedom (Kuznetsova et al., 2017). Participant's patterns of sorting behavior were characterized by calculating the Euclidean distance between images, which were then normalized for each participant by scaling distances based on the maximum distance for each participant. I first conducted a series of analyses using top-down, supervised approaches, followed by a series of analyses using bottom-up, unsupervised approaches. The analyses conducted were exploratory in nature, implementing similar approaches to those applied in past studies using the spatial arrangement method (Unger et al., 2016); however, converging patterns of results across multiple different analyses give me increased confidence in the robustness results of this study. For an analysis of possible gender differences in the task see Appendix B.

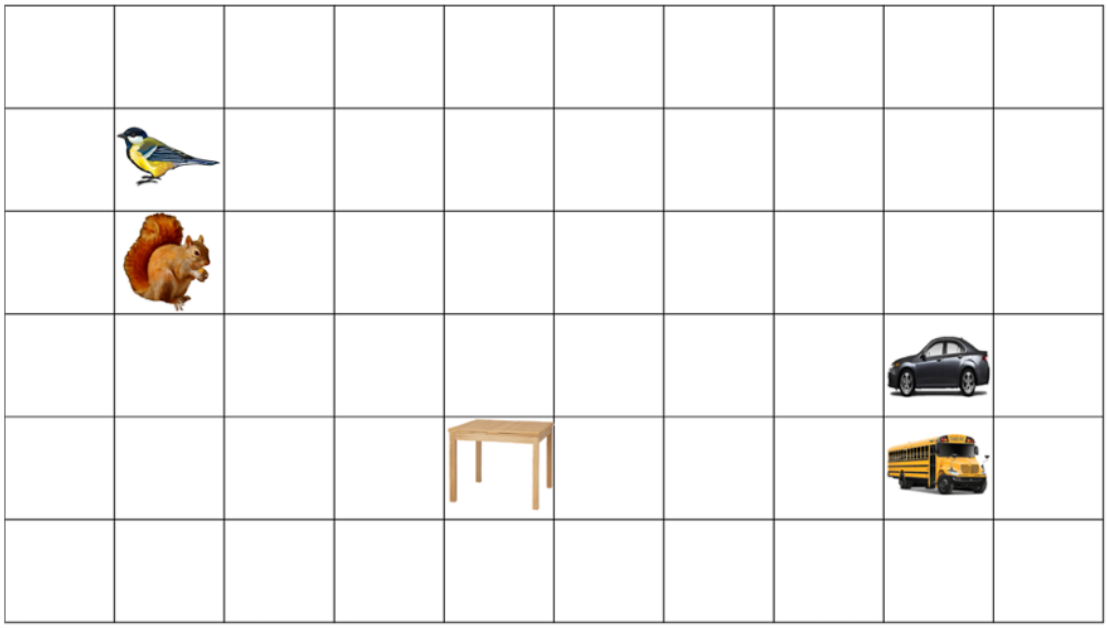
Practice Phase

To assess whether participants in each age group demonstrated understanding of SpAM during the practice phase, I investigated the degree to which participants consistently arranged images belonging to the same superordinate categories (vehicles: car and bus; animals: squirrel and bird) closer together in space (Figure 3). I computed the average distance between images

belonging to the same superordinate category (vehicles or animals) for each participant, and then compared the average distances for item pairs sharing the same category to item pairs from different categories. 3-year-olds did not consistently arrange items belonging to the same superordinate category closer together in space, (paired t-test: $t(20) = -0.046$, $p = .96$).

Figure 3

Example of a Participant's Practice Sort



This suggests that children in this age group were not consistently sorting images according to superordinate categories and may have struggled with the task instructions. However, I included 3-year-olds in all analyses as there was no reason to exclude this group a priori. Furthermore, the inclusion or exclusion of 3-year-olds does not alter any of the patterns of results in the study. All other age groups consistently sorted images belonging to the same category closer together in the grid space (4-year-olds: $t(34) = 5.03$, $p < .001$; 5-year-olds: $t(27) = 5.42$, $p < .001$; 6-year-olds: $t(22) = 8.55$, $p < .001$; adults: $t(39) = 29.10$, $p < .001$).

Comparing Dimensions of Affect and Categories in Sorting Behaviors.

Next, I used top-down, supervised methods to examine whether emotion category and dimensions of affect account for how closely different facial cues are placed to one another. I examined these features separately, and then compare how well the various dimensions and categories account for sorting behaviors.

Emotion Category

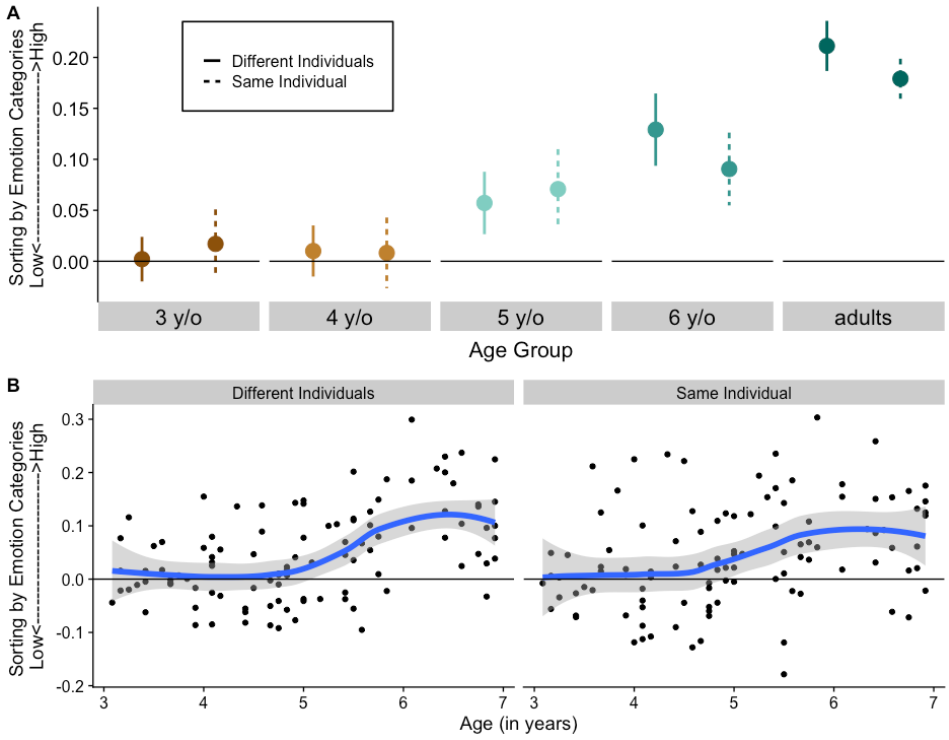
I first investigated developmental change in the use of common English language emotion categories (e.g., sad, happy, anger, disgust, fear, surprise, neutral, calm, excitement) as a structure for emotion cues. To do so, I computed the average distance between images that shared the same category label (e.g., the distance of one happy face to another happy face) versus images that had differing category labels (e.g., the distance of one happy face to a sad face) for each participant (see also Unger et al., 2016 for a similar approach). To do so, I fit a linear mixed-effects model estimating the average distance between item pairs for adults versus children (coded .5,-.5), the category match for an image pair (same category pair vs. different category pair; centered: same = 0.5, different = -0.5), and their interaction with a by-participant random intercept and a by-participant random slope for category match. I analyzed results collapsing across sorting conditions, as there was no evidence that results differed between the same and different individual sorts (p = .28). Adults were more likely than children to place images belonging to the same emotion categories closer together than images belonging to different emotion categories, b = -0.15, Wald 95% CI = [-0.18,-0.12], $F(1,173.40) = 125.90$, $p < 0.001$.

To understand how children's use of emotion categories changed across development, I next fit a linear mixed-effects model on the child data with age (in years; centered) as a

continuous predictor with an otherwise identical model structure. Children were more likely to sort facial configurations based upon emotion category labels with increasing age, b = -0.03, Wald 95% CI = [-0.05, -0.02], $F(1,130.91) = 28.13$, $p < 0.001$. This developmental increase in use of category labels is shown in Figure 4. Follow-up analyses of each age group separately reveals that neither 3-year-olds (p = .45) nor 4-year-olds (p = .47) showed evidence of sorting based upon emotion categories, while 5-year-olds (b = -0.06, Wald 95% CI = [-0.09, -0.04], $F(1,31.81) = 22.27$, $p < .001$) and 6-year-olds (b = -0.11, Wald 95% CI = [-0.14, -0.08], $F(1,33.94) = 65.47$, $p < .001$) began using category information, though to a lesser extent than adults (b = -0.20, Wald 95% CI = [-0.21, -0.18], $F(1,69.99) = 488.21$, $p < .001$).

Figure 4

Use of Emotion Category When Sorting Facial Cues



Note. Use of emotion categories in sorting behavior by A) age-bin and B) continuous age. Y-axis represents the difference in average distance for items belonging to the same vs. different emotion categories. An average value of zero represents no distinction by emotion category, as faces from the same versus different emotion categories were equally far apart. Error bars represent 95% confidence intervals.

To investigate the strength of these results, I also investigated the consistency with which children used emotion category across different emotions and sorts. Similar developmental patterns in the use of emotion category occurred across all categories (Figure 5). To better assess these similarities, I also examined correlations in the use of categories across the two sorts. First, I correlated the average distances at which category-based item pairs (happy versus sad) were placed across the two sorts (Table 1). Second, I investigated these correlations at the participant level using a repeated measures correlation (Table 2). Across both analyses, category-based correlations between sorting conditions were small for 3- and 4-year-olds and became increasingly robust with the highest correlations among adults. This pattern mirrors the earlier analyses, with 3- and 4-year-olds showing weak or no evidence for category-based sorting and increasing evidence for category-based sorting from the age of 5 onward.

Table 1

Correlations for Category-Based Sorting Distances Across Sorting Conditions

Age Groups	***Correlation***	***95% CI***	***t-value***	***df***	***p-value***
3-year-olds	0.39	[0.11, 0.61]	2.75	43	.009
4-year-olds	0.34	[0.05, 0.57]	2.34	43	.024
5-year-olds	0.60	[0.37, 0.76]	4.88	43	<.001
6-year-olds	0.67	[0.47, 0.81]	5.92	43	<.001
adults	0.67	[0.46, 0.8]	5.87	43	<.001

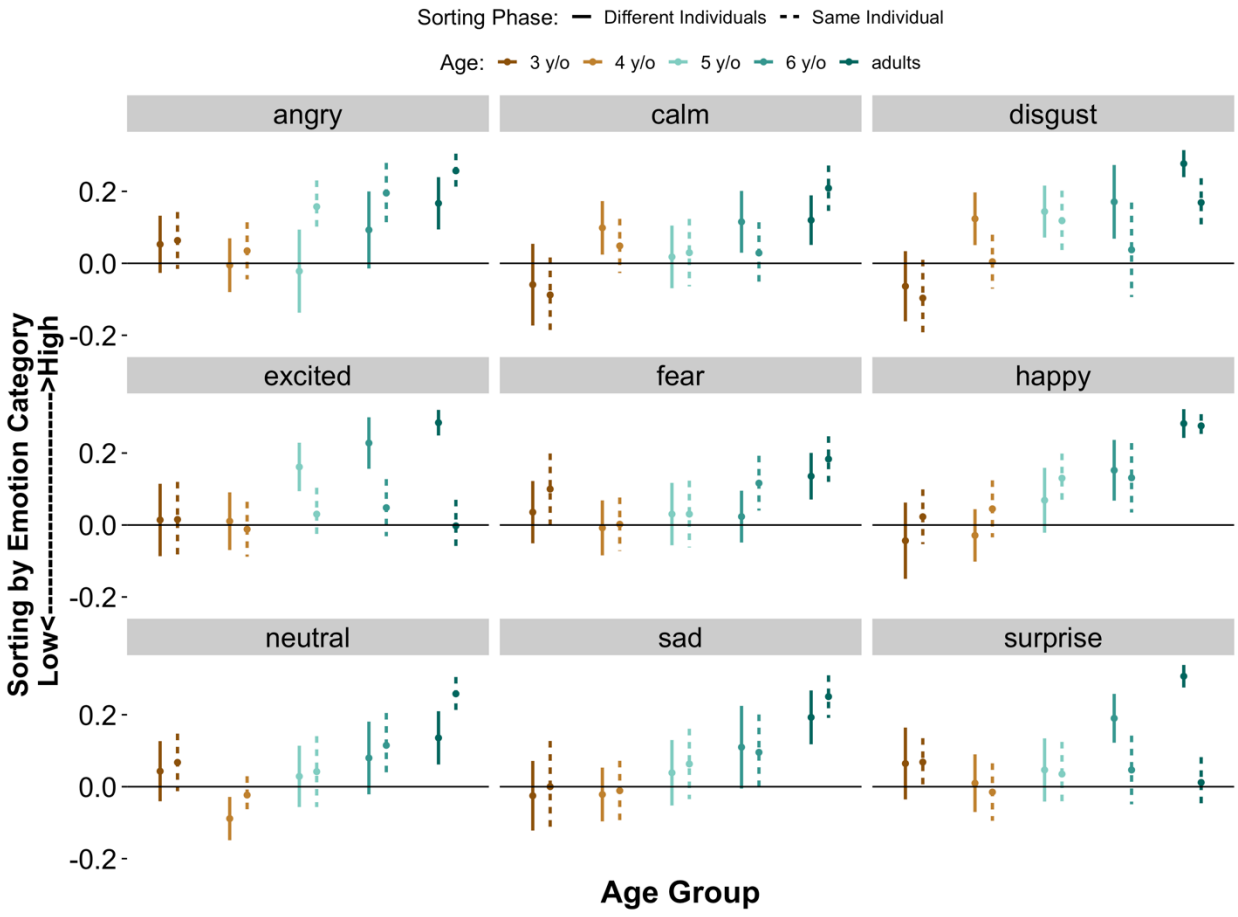
Table 2

Repeated Measures Correlations for Category-Based Sorting Across Sorting Conditions

Age Groups	*Correlation*	*95% CI*	*df*	*p-value*
3-year-olds	0.08	[0.01, 0.14]	923	.016
4-year-olds	0.03	[-0.02, 0.08]	1451	.30
5-year-olds	0.12	[0.07, 0.18]	1231	<.001
6-year-olds	0.22	[0.16, 0.28]	1011	<.001
adults	0.36	[0.32, 0.4]	1715	<.001

Figure 5

Use of Emotion Categories in Sorting Behavior by Age and Emotion Category



Note. Y-axis represents the difference in average distance for items belonging to the same vs. different emotion categories. An average value of zero represents no distinction by emotion category, as faces from the same versus different emotion categories were equally far apart. Error bars represent 95% confidence intervals.

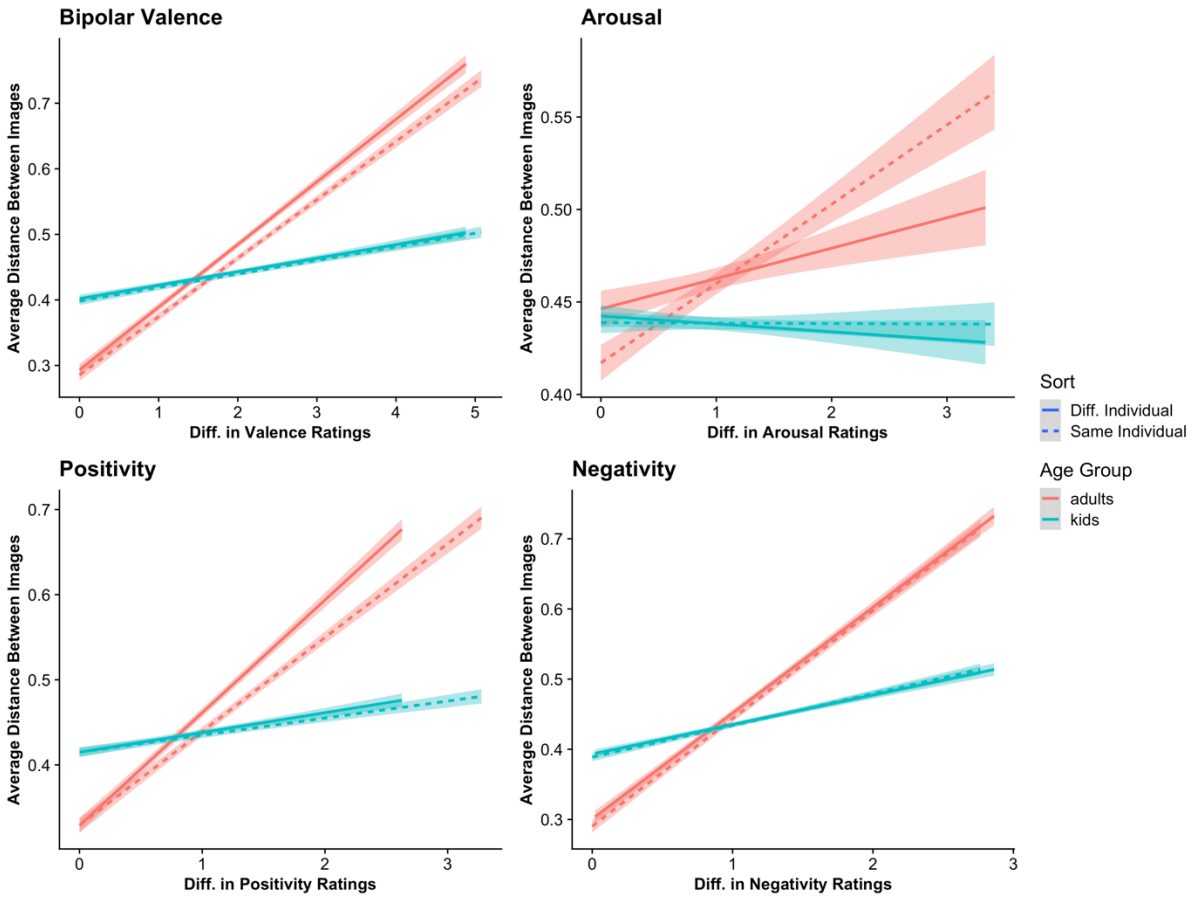
Dimensions of Affect

Next, I tested whether bipolar valence, bivariate valence (separate ratings of positivity and negativity), and arousal predicted participants' sorting behavior. To do so, I fit a series of linear mixed-effects models regressing the average distance between item pairs on their similarity along the dimension of interest (bipolar valence, arousal, positivity, and negativity) – measured in terms of the difference in average stimulus rating between image pairs. This analysis included age group (adults: .5; children: -.5), its interaction with the dimension of interest, and random effects for items and participants, including a by-participant random intercept, a by-participant random slope for the dimension of interest, and a by-item random intercept. Adults were more likely than children to use each of the four dimensions to guide their sorting behaviors (bipolar valence: $b = 0.07$, Wald 95% CI =[0.06, 0.08], $F(1, 145.47) = 148.52$, $p < .001$; arousal: $b = 0.03$, Wald 95% CI =[0.02, 0.04], $F(1, 143.22) = 41.96$, $p < .001$; positivity: $b = 0.10$, Wald 95% CI =[0.08, 0.11], $F(1, 145.56) = 146.66$, $p < .001$; negativity: $b = 0.11$, Wald 95% CI =[0.09, 0.13] $F(1, 145.38) = 112.14$, $p < .001$; see Figure 6). To further understand the developmental change in children's use of each dimension, I fit linear mixed-effects models on the child data with age (in years; centered) as a continuous predictor and an otherwise identical model structure. Children increasingly used each feature across development (valence: $b = 0.01$, Wald 95% CI =[0.01, 0.02], $F(1, 103.89) = 31.08$, $p < .001$; positivity: $b =$

0.01, Wald 95% CI =[0.01, 0.02], $F(1, 103.95) = 21.34$, $p < .001$; negativity: $b = 0.03$, Wald 95% CI =[0.02, 0.03], $F(1, 103.79) = 35.35$, $p < .001$)— with the exception of arousal, $b = 0.003$, Wald 95% CI =[-0.001, 0.01], $F(1, 103.81) = 1.97$, $p = .16$. The pattern for arousal highlights how children's development may not always occur as straightforward linear differentiation.

Figure 6

Use of Dimensions in Sorting Behavior by Sort and Age Group



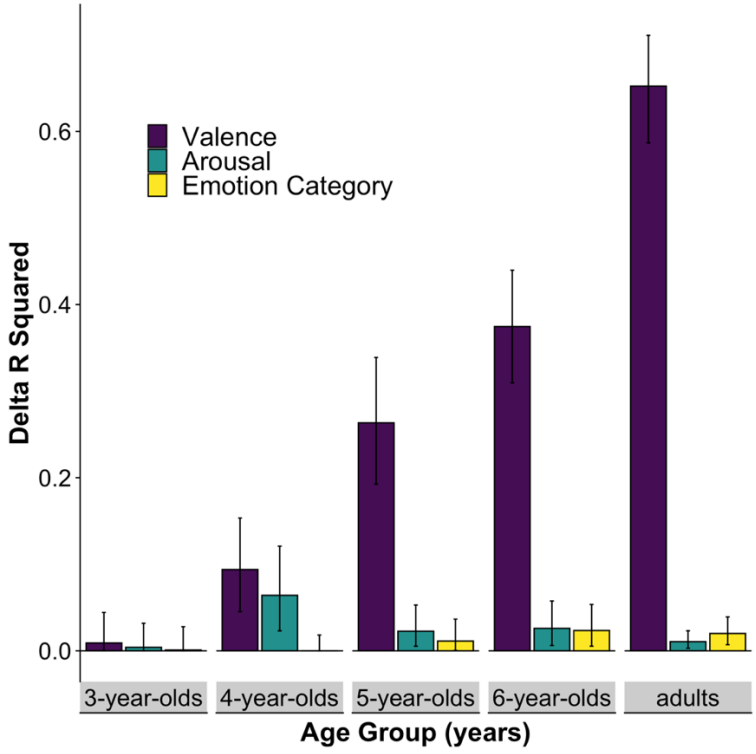
Note. Use of dimensions (valence, arousal) in sorting behavior by sort and age group. Y-axis represents the average distance between an item pair (e.g., happy-sad). X-axis represents the average rating difference between an item pair. Shading represents 95% confidence intervals. A positively increasing slope indicates that images with similar ratings are placed closer together.

Comparing Dimensions of Affect and Emotion Categories

Finally, I examined how well emotion category predicted participant's sorting behavior compared to valence and arousal. To do so, I computed the average distance between all stimulus pairs (n = 306 unique pairs) for each age group and predicted these distances from a pair's similarity on each dimension of interest simultaneously. This general linear model revealed how much each dimension aided in explaining variance in each age group's sorting behavior. First, I estimated the use of bipolar valence, arousal, and whether image pairs shared the same discrete emotion category (0 = different category pair; 1 = same category pair). Second, I estimated the effects of bivariate valence with positivity and negativity as two orthogonal dimensions.

Figure 7

Delta R-squared for each Predictor of Sorting Behavior



Note. Error bars represent bootstrapped 95% confidence intervals.

Bipolar Valence, Arousal, and Shared Emotion Category. Valence emerged as (by far) the strongest predictor (Figure 7) of how participants grouped facial images, an effect that increased steadily with age. Arousal was a significant predictor for 4-year-olds but declined as children grew older. Consistent with the results from the previous section, emotion category did not emerge as a predictor until age 5 years. The total variance explained by this model increased steadily across age (Table 3), accounting for a significant amount of the error variance for all age groups: $F(3, 302) > 14$, $p < .001$) with the exception of the youngest age group (3-year-olds: $F(3, 302) = 1.33$, $p = .26$.

Table 3

Predicting Sorting Distance from Valence, Arousal, and Shared Emotion Category

Predictor	Estimate	*t*-value	*p*	ΔR^2	Overall R^2
3-year-olds					.01
Valence	0.004	1.67	.10	.01	
Arousal	-0.004	-1.12	.27	.00	
Emotion Category	-0.006	-0.55	.58	.00	
4-year-olds					.13
Valence* **	0.01	5.71	<.001	.09	
Arousal* **	-0.02	-4.71	<.001	.06	
Emotion Category	-0.001	-0.15	.88	.00	
5-year-olds					.31
Valence* **	0.03	10.76	<.001	.26	
Arousal**	-0.02	-3.16	.002	.02	
Emotion Category*	-0.03	-2.22	.027	.01	
6-year-olds					.46
Valence* **	0.05	14.49	<.001	.37	
Arousal* **	-0.02	-3.82	<.001	.03	
Emotion Category* **	-0.06	-3.63	<.001	.02	
Adults					.78
Valence* **	0.09	30.04	<.001	.65	
Arousal* **	-0.02	-3.83	<.001	.01	
Emotion Category* **	-0.08	-5.26	<.001	.02	

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

Bivariate Valence, Arousal, and Shared Emotion Category. I repeated the previous analysis, replacing bipolar valence with bivariate valence (positivity and negativity). As expected, ratings of positivity and negativity were highly correlated with bipolar ratings, precluding me from including all five predictors in the same model. The dimension of negativity emerged as the strongest predictor of sorting behavior across age ranges, even 3-year-olds, and explained substantially more variance than positivity, arousal, and emotion category (Table 4).

Table 4

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

Predictor	Estimate	*t*-value	*p*	ΔR^2	Overall R^2
3-year-olds					.02
Positivity	-0.005	-1.01	.31	.00	
Negativity*	0.01	2.23	.027	.02	
Arousal	-0.00004	-0.01	.99	.00	
Emotion Category	-0.004	-0.33	.74	.00	
4-year-olds					.18
Positivity	-0.003	-0.66	.51	.00	
Negativity***	0.02	5.60	<.001	.09	
Arousal*	-0.009	-2.42	.02	.02	
Emotion Category	0.004	.42	.67	.00	
5-year-olds					.45
Positivity*	-0.01	-2.48	.014	.01	
Negativity***	0.07	12.02	<.001	.26	
Arousal	0.006	1.22	.22	.00	
Emotion Category	-0.02	-1.31	.19	.00	
6-year-olds					.57
Positivity	-0.01	-1.27	.21	.00	
Negativity***	0.09	13.87	<.001	.27	
Arousal	0.01	0.92	.36	.00	
Emotion Category**	-0.04	-2.89	.004	.01	
Adults					.80
Positivity***	0.05	8.36	<.001	.05	
Negativity***	0.11	15.39	<.001	.16	
Arousal	-0.003	-0.54	.59	.00	
Emotion Category***	-0.08	-4.92	<.001	.02	

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

Does Bivariate Valence Predict Sorting Behavior Better than Bipolar Valence? To determine whether separate dimensions of positivity and negativity were better predictors than bipolar valence, I compared the models including bipolar valence to the models including positivity and negativity (bivariate valence) in each age group. Bivariate dimensions of valence were a better predictor of sorting behavior in all but the youngest age group, with the most substantial gains among the 5- and 6-year-olds (3-year-olds: $F(1, 301) = 2.82$, $p = .09$; 4-year-olds: $F(1, 301) = 18.98$, $p < .001$; 5-year-olds: $F(1, 301) = 74.58$, $p < .001$; 6-year-olds: $F(1, 301) = 77.51$, $p < .001$; adults: $F(1, 301) = 21.88$, $p < .001$).

Bottom-up Assessment of Facial Cues of Emotion

Next, I conducted a series of analyses using unsupervised methods to provide a complementary perspective on how emotions might be represented. The Same and Individual Sorts were analyzed separately because the following analyses require pairwise distances between all items, which are only available within a given sorting block. The unsupervised analyses extract patterns from the sorting data by using the pairwise distances between all of the stimuli without regard to the labels or affective ratings of those stimuli. This allows me to represent differences in how children and adults are approaching the task without relying on any predetermined dimensions or categories. In order to facilitate comparisons between all of the analyses in the paper, I also investigate the extent to which sorting patterns extracted in the unsupervised analyses can be predicted from emotion category labels and affective dimensions.

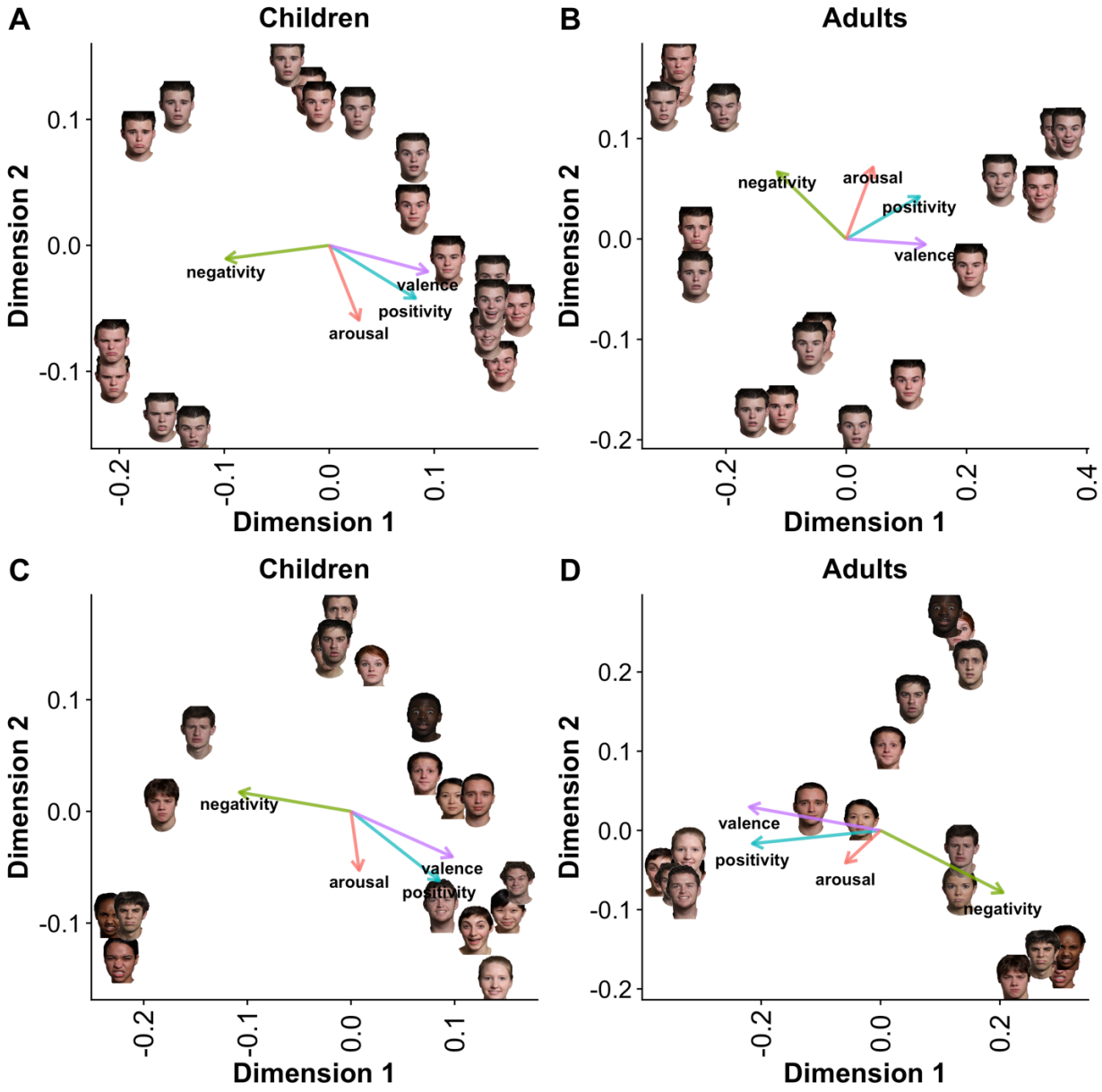
Multidimensional Scaling

First, I used 2-dimensional classical multidimensional scaling (MDS) to visually represent participants' sorting behaviors (Figure 8). MDS uses a dissimilarity matrix to make an n-dimension representation of the data in which the dimensions provide information about

underlying relationships in the data (Hout, Goldinger, et al., 2013). I found that the 2-dimensional MDS solution had the best fit based on the elbow of an eigenvalue by n-dimensions plot. The 2-dimensional structure also seemed most appropriate given the nature of the task (e.g., sorting all images together in a 2-dimensional space), and given the proportion of variance explained by the MDS solutions for adults (Same Individual: $R^2 = 0.67$, Different Individuals: $R^2 = 0.68$) and children (Same Individual: $R^2 = 0.30$; Different Individuals: $R^2 = 0.32$). Because interpretation of MDS dimensions can be subjective (Hout, Papesh, et al., 2013), I fit vectors of image ratings for bipolar valence, arousal, and bivariate valence (positivity and negativity) onto the MDS solution over 1,000 permutations to derive the squared correlation coefficient of each vector (envfit in the R package vegan; Oksanen, 2019). This analysis revealed that stimuli ratings of bipolar valence, positivity, and negativity consistently correlate with the MDS dimensions ($r^2 > 0.84$ and $p < .001$ across all sort conditions for both adults and children). Arousal only correlated with the dimensions in the Same Individual Sort (Adults: $r^2 = 0.40$, $p < .05$; Children: $r^2 = 0.43$, $p < .05$) and not in the Different Individual Sort (Adults: $r^2 = 0.10$, $p = .46$; Children: $r^2 = 0.22$, $p = .16$). To give more insight into developmental trajectories underlying MDS dimensions, I also looked at MDS by age bin (Table 5). I again found that 3-year-olds' sorting behavior was not well captured, although the dimensions of positivity and negativity were marginally significant in the Different Individuals Sort. Older children's sorting was best captured by valence—and sometimes arousal.

Figure 8

Classical Multidimensional Scaling Solution (2-dimensions) Across All Children and Adults in the Same Individual (A, B) and Different Individual (C, D) Sorts



Notes. Vectors show squared correlation coefficients between image ratings and the MDS dimensions.

Table 5

Squared Correlation Coefficients for Classical MDS Solutions

Age Group	***Same Individual Sort***				***Different Individuals Sort***			
	Pos.	*Neg.*	*Valence*	*Arousal*	*Pos.*	*Neg.*	*Valence*	*Arousal*
Adults	**0.92*** **	**0.98*** **	**0.96*** **	**0.40***	**0.92*** **	**0.96*** **	**0.98*** **	0.10
Children	**0.84*** **	**0.97*** **	**0.92*** **	**0.43***	**0.85*** **	**0.91*** **	**0.86*** **	0.22
3-year-olds	0.02	0.04	0.2	0.2	0.23	0.29†	0.30†	0.0
4-year-olds	**0.65*** **	**0.85*** **	**0.79*** **	0.08	**0.47*** **	**0.66*** **	**0.55***	0.0
5-year-olds	**0.76*** **	**0.96*** **	**0.90*** **	0.18	**0.64*** **	**0.73*** **	**0.65*** **	**0.32***
6-year-olds	**0.86*** **	**0.92*** **	**0.89*** **	**0.42***	**0.87*** **	**0.93*** **	**0.90*** **	0.25

Note. Asterisks denote significance level, *$p < .05$; **$p < .01$; ***$p < .001$, $p < 0.1$ = '†'

In addition to classical multidimensional scaling, I also examined non-metric multidimensional scaling (NMDS). Classical MDS assumes the distance measured is exact (as when measuring distances on a map), while NMDS allows for more error and focuses on the ranking of the distances rather than the exact distance itself. For this reason, NMDS may be more appropriate for more subjective similarity ratings (Wickelmaier, 2003). The 2-dimensional NMDS solution had an acceptable fit, as the stress values ranged from excellent, *Stress 1* <5%, to fair, *Stress 1*<14% (Kruskal, 1964; Kruskal's stress (*Stress 1)* for Same Individual: Adult = 5.40%, Children = 13.65%; Different Individuals: Adult = 4.33%, Children = 12.61%). Similar to the classical MDS results, valence related to the dimensions, while arousal often did not (Table 6). Across both classical and non-metric MDS, there was not strong and consistent support for bipolar valence and arousal serving as underlying dimensions in the task. Rather, bivariate valence (positivity and negativity) best captured the underlying dimensions.

Table 6

Squared Correlation Coefficients for Non-metric MDS Solutions

Age Group	***Same Individual Sort***				***Different Individuals Sort***			
	Pos.	*Neg.*	*Valence*	*Arousal*	*Pos.*	*Neg.*	*Valence*	*Arousal*
Adults	**0.91*** **	**0.96*** **	**0.97*** **	0.27†	**0.89*** **	**0.96*** **	**0.98*** **	0.07
Children	**0.73*** **	**0.93*** **	**0.86*** **	0.20	**0.79*** **	**0.83*** **	**0.78*** **	0.17
3-year-olds	0.04	0.10	0.05	0.01	0.27†	0.28†	**0.35***	0.03
4-year-olds	**0.61****	**0.75*** **	**0.71*** **	0.10	**0.43***	**0.57*** **	**0.44***	0.05
5-year-olds	**0.72*** **	**0.88*** **	**0.84*** **	0.12	**0.57****	**0.60****	**0.54****	0.28†
6-year-olds	**0.74*** **	**0.87*** **	**0.80*** **	**0.40***	**0.84*** **	**0.87*** **	**0.86*** **	0.26

Note. Asterisks denote significance level, *$p < .05$; **$p < .01$; ***$p < .001$, $p < 0.1$ = '†'

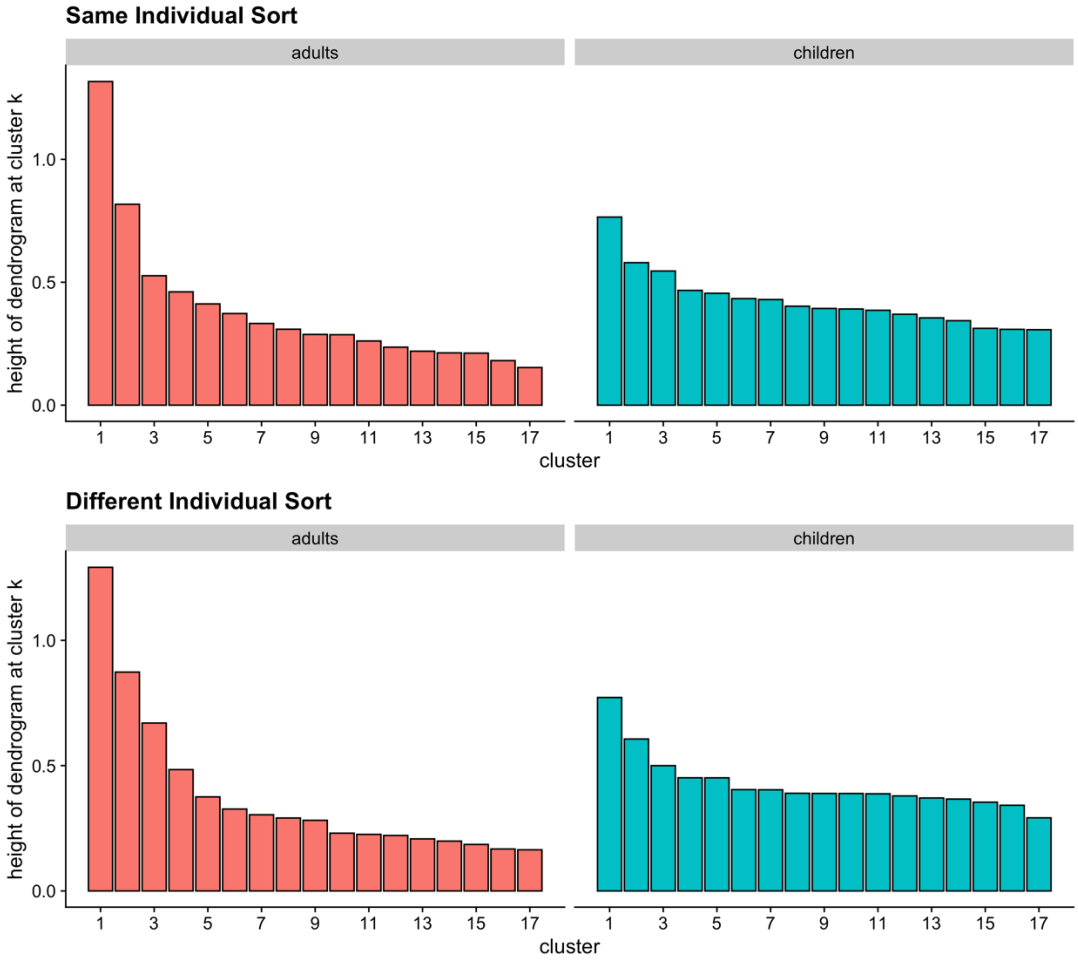
Hierarchical Clustering

Next, I used hierarchical clustering to examine age-related changes in how participants organized emotion cues, as in prior work with children and SpAM (e.g., Unger et al., 2016, Vales, Stevens, et al., 2020). This analysis allowed us to examine similarities in how adults and children sorted the facial stimuli, without using adult emotion categories or affective ratings to represent similarity between stimuli. To select the hierarchical clustering method (average, single, complete, or Ward's), I analyzed the agglomerative coefficient, a measure of the strength of the clustering structure. I selected Ward's method (Ward, 1963) because it had the highest agglomerative coefficient across all age bins and sorting conditions, and is a generally preferred method for agglomerative hierarchical clustering (Boehmke & Greenwell, 2020; Kaufman & Rousseeuw, 2009). Clustering was performed on distance matrices calculated for each age group in each sorting condition using the pairwise distances between all sorted images. I opted to highlight the three cluster (k=3) solution in many of the analyses below after examining changes in the height of the dendrogram at different clusters (Figure 9), subjective evaluation of the dendrograms, patterns of valence across the different clusters, and a desire to be consistent across

different age ranges. However, all possible clusters (k=1-18) are still visible in the dendrograms (Figure 10), and I report similarity indices for multiple values of k.

Figure 9

Bar Plots of the Height of Children and Adults' Dendrograms at Different Cluster Sizes



Note. Examining changes in these heights can be helpful for determining the optimal number of clusters to display as height is an indicator of similarity.

I used a number of indices of similarity to examine changes in hierarchical clustering (represented by dendrograms) across development. To examine the similarity of the dendrograms

as a whole, I used Baker's gamma, the cophenetic correlation coefficient, and entanglement (Table 7). To examine the similarity of particular clusters within the dendrograms (e.g., k=3) I used the Adjusted Rand index and the Fowlkes-Mallows (FM) Index (Table 8). Across all indices, I found the same general pattern of children's clustering structures becoming increasingly adult-like. These changes appear to be driven by changes in emotion knowledge and not improvement on the task generally, as the practice structure is highly similar to adults for all age groups except 3-year-olds. Changes in children's clusters otherwise show strong evidence of systematicity, as children closer in age are more similar to one another (see correlations in Figure 10). For example, the sorting behavior of 5-year-olds had a stronger correlation with 6-year-olds and 4-year-olds than with adults.

Table 7

Comparison of Children's and Adult's Dendrograms

	Practice			***Same Individual***			***Different Individuals***		
Age Group	*Baker's Gamma*	*c*	*Ent.*	*Baker's Gamma*	*c*	*Ent.*	*Baker's Gamma*	*c*	*Ent.*
3-year-olds	-.19	-.03	0	-.01	.0	.33	.17	.21	.23
4-year-olds	.5	.86	0	.16	.2	.15	.18	.2	.16
5-year-olds	1.0	.99	0	.43	.41	.10	.36	.38	.08
6-year-olds	1.0	.98	0	.51	.65	.07	.30	.40	.09

Note. Each value in the table represents the similarity between children's and adults' dendrograms for different age bins. Both the cophenetic correlation *c* and the Baker's Gamma Index range from -1 to 1 with values near 0 suggesting that the two dendrograms are not statistically similar (Baker, 1974; Sokal & Rohlf, 1962). Entanglement (Ent.) measures whether the labels at the bottom of two dendrograms match one another. A value of 0 indicates perfect alignment, while a value of 1 indicates no alignment. To optimize alignment, I first used the untangle method (Galili, 2015).

Table 8

Comparison of Children's and Adult's Clusters (k=3-5)

	Practice		***Same Individual***						***Different Individuals***					
Similarity Index	*Adj. Rand*	*FM*	*Adjusted Rand*			*FM*			*Adjusted Rand*			*FM*		
Clusters (k)	**3**	**3**	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**
3-year-olds	.21	.41	.02	.10	.03	.32	.31	.21	.16	.12	.12	.44	.32	.28
4-year-olds	1.0	1.0	.14	.39	.35	.40	.54	.47	.14	.11	.10	.42	.32	.26
5-year-olds	1.0	1.0	.49	.31	.45	.67	.48	.54	.49	.38	.37	.65	.53	.48
6-year-olds	1.0	1.0	.83	.66	.50	.88	.75	.60	.38	.46	.52	.58	.60	.62

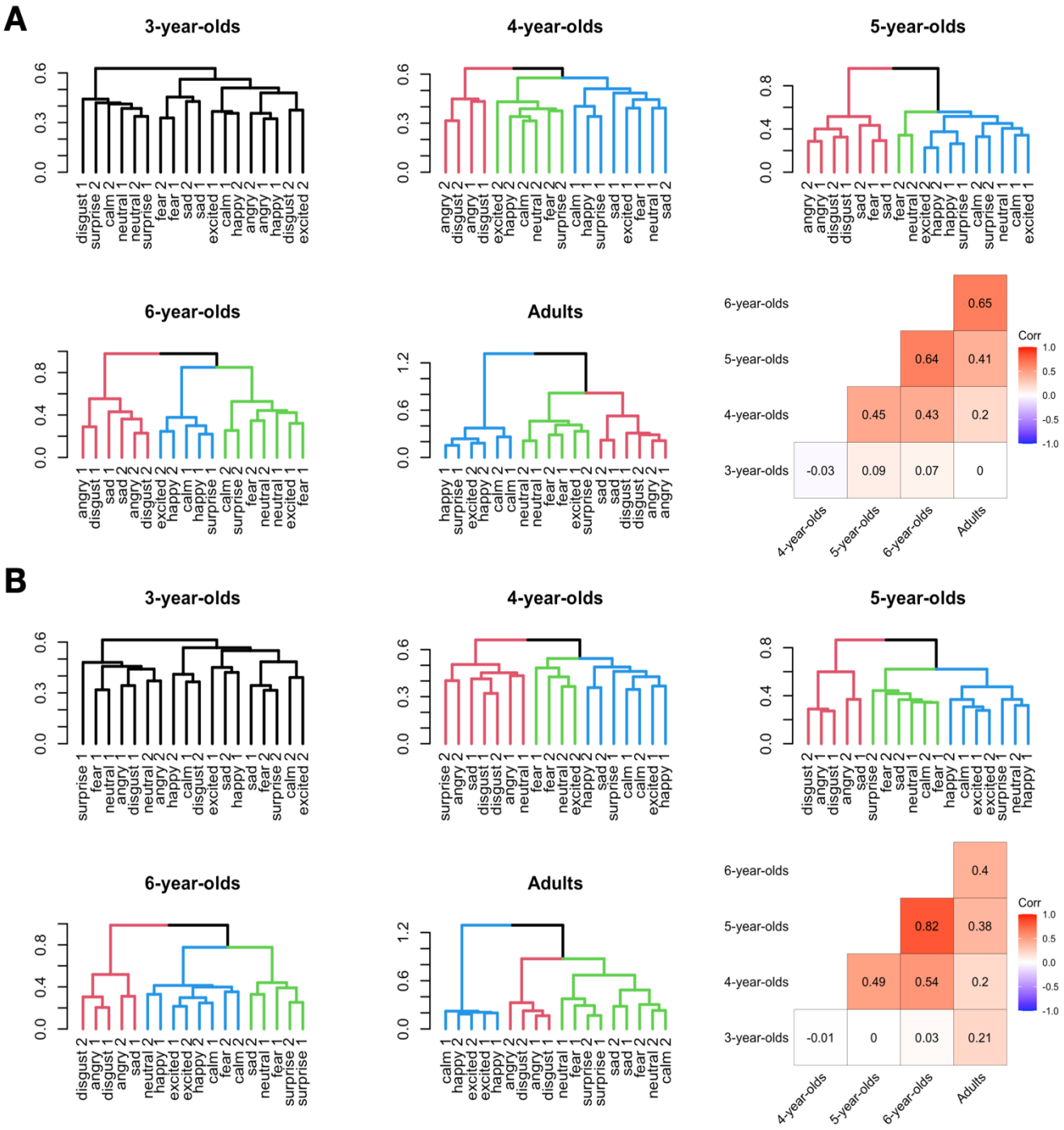
Note. Each value in the table represents the similarity between children's clusters for a particular value of *k* (3,4,5) and adults' clusters. The Adjusted Rand Index measures the likelihood that the same items appear in the same clusters, while controlling for the fact that this will sometimes happen due to random chance. An adjusted Rand Index near 0 indicates two clusters match no more than expected by random chance, with higher and lower values indicating higher- or lower-than-chance-level similarity between the two clusters (Hubert & Arabie, 1985; Rand, 1971). The Fowlkes-Mallows (FM) Index also measures the similarity between two clusters (Fowlkes & Mallow, 1983). It ranges from 0 (no similarity) to 1 (perfect similarity) and aims to capture how likely it is that two clusters contain the same items.

These changes in organizational structure can also be seen in dendrograms, which visualize the hierarchical clustering solutions. Each facial image is a node on the dendrogram that forms another node (represented by a horizontal line) when it merges with another face. Clusters are determined by the vertical height of the branches in a dendrogram, not by which labels are closest to one another laterally. Thus, faces that were found to be the most similar would be connected as a node with a very low height. The three-cluster solution highlighted in

the dendrograms was strongly related to the valence ratings of the images. To examine this relation, I ran a linear regression model predicting the bipolar valence ratings of the facial stimuli using cluster group (k = 3). As in the distance-based analyses above, bipolar valence was a strong predictor of both children's (Same Individual Sort: $F(2,15)=73.86$, $p<.001$; Different Individual Sort: $F(2,15)=26.03$, $p<.001$), and adults' (Same Individual Sort: $F(2,15)=51.87$, $p<.001$; Different Individual Sort: $F(2,15)=61.71$, $p<.001$) cluster groups for all age bins except for 3-year-olds (Same Individual Sort: $F(2,15)=1.48$, $p=.26$; Different Individual Sort: $F(2,15)=2.63$, $p=.10$).

Figure 10

Dendrograms and Correlations between Dendrogram Structures for the (A) Same Individual Sort and (B) Different Individual Sort



Note. The numbers 1 and 2 indicate that the images had open and closed mouths, respectively. Colors specify the three cluster solutions for each age bin and highlight commonalities across dendrograms. Red clusters contain mainly anger and disgust images, green clusters mainly contain certain fear and neutral images, and blue images mainly contain certain happy, calm, and surprise images. 3-year-olds' dendrograms are colored differently as they showed less differentiation.

Verbal Fluency Task

Children's responses during the verbal fluency task were recorded and coded by two independent coders. The two coders agreed 93.87% of the time (95.98% for emotion words and 93.75% for animal words). A third coder rated all participants with discrepancies a third time, and then resolved all discrepancies between the other two coders. Most of the differences involved either coders reporting similar but different words (e.g., frog versus dog), one coder recording a word that another coder did not (e.g., one coder reports "sad" and the other did not), or one coder recorded a repetition that another did not (e.g., a child said "sad" followed by "sad" again). In one instance, a coder paired the wrong audio file with the participant.

Inclusion criteria for emotion and animal words were quite liberal. For animals, fantasy (e.g., dragon) and extinct animals (t-rex) were included. Additionally, different animal species (e.g., black bear, polar bear, grizzly bear) were treated as different tokens. Terms indicating age and size did not count as unique tokens (e.g., baby pig, little pig, and pig would all count as "pig"), and plants and fantasy human creatures (e.g., witches) were excluded (n = 2). For emotion words, terms indicating different amounts of feeling did not count as unique tokens (e.g., a little happy, kind of happy, and very happy would all count as "happy"), and word forms

were standardized across participants during analysis (e.g., “disgust”, “disgusting”, and “disgusted” were all treated as “disgust” and not as three separate terms). All duplicates within participants were excluded (n = 58) for a total of 1,121 trials.

Descriptive Analysis

On average, children produced 4.7 more animal than emotion words (paired t-test: $t(89) = 10.29$, $p < .001$), and older children (age continuous, centered) produced more animal ($b = 2.24$, $SE = 0.44$, $t = 5.07$, $p < .001$) and emotion words overall ($b = 1.22$, $SE = 0.19$, $t = 6.44$, $p < .001$, see Table 9). Children also displayed greater consensus in the emotion words produced (see Figure 11). For instance, sad (93.3%) and happy (89.3%) were produced by nearly all participants in the emotion word condition, while the most common animal word was only produced by half of participants (lion, 51.11%). This was also reflected in the variety of words produced as there were more unique animal tokens (n=147) than emotion tokens (n=54).

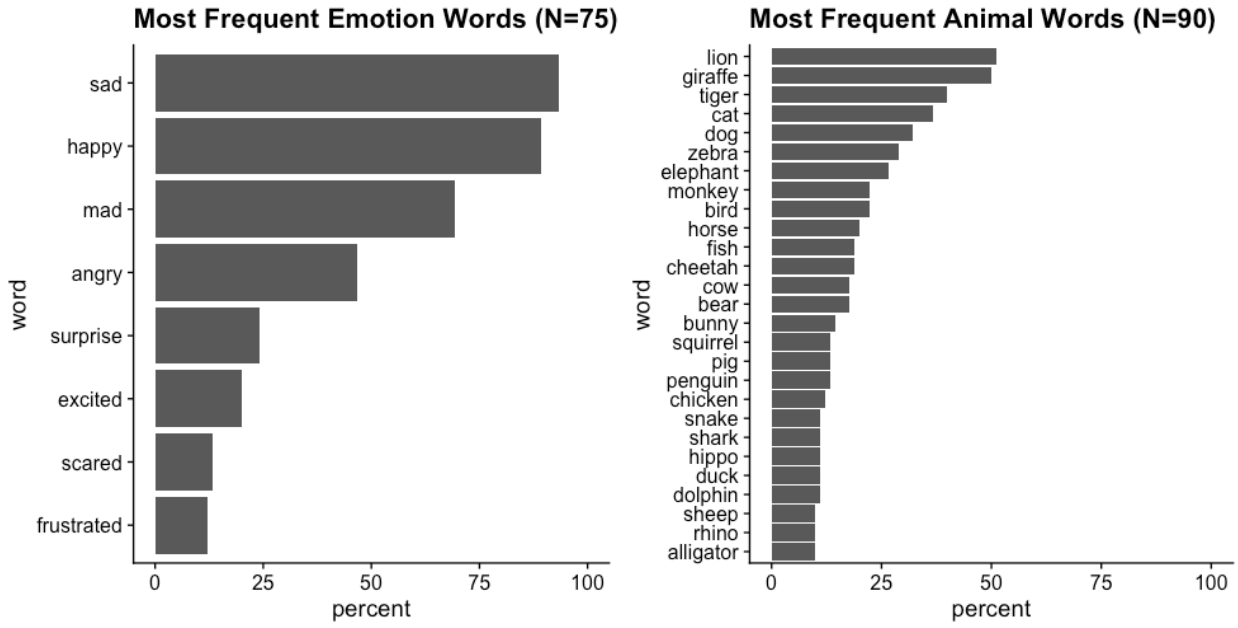
Table 9

Average Word Count by Age Bin and Condition (N=90)

Age Group	***Animal Words***		***Emotion Words***	
	Average	*Range*	*Average*	*Range*
Overall	8.58	1-24	3.88	0-9
3-year-olds	5.23	1-11	1.31	0-6
4-year-olds	6.73	1-15	3.5	0-8
5-year-olds	9.65	3-22	4.42	0-7
6-year-olds	12.0	1-24	5.33	2-9

Figure 11

All Emotion and Animal Words Produced by at Least 10% of Participants



Note. All emotion and animal words produced by at least 10% of participants. Percentages were calculated using participants that stated at least one word from that category.

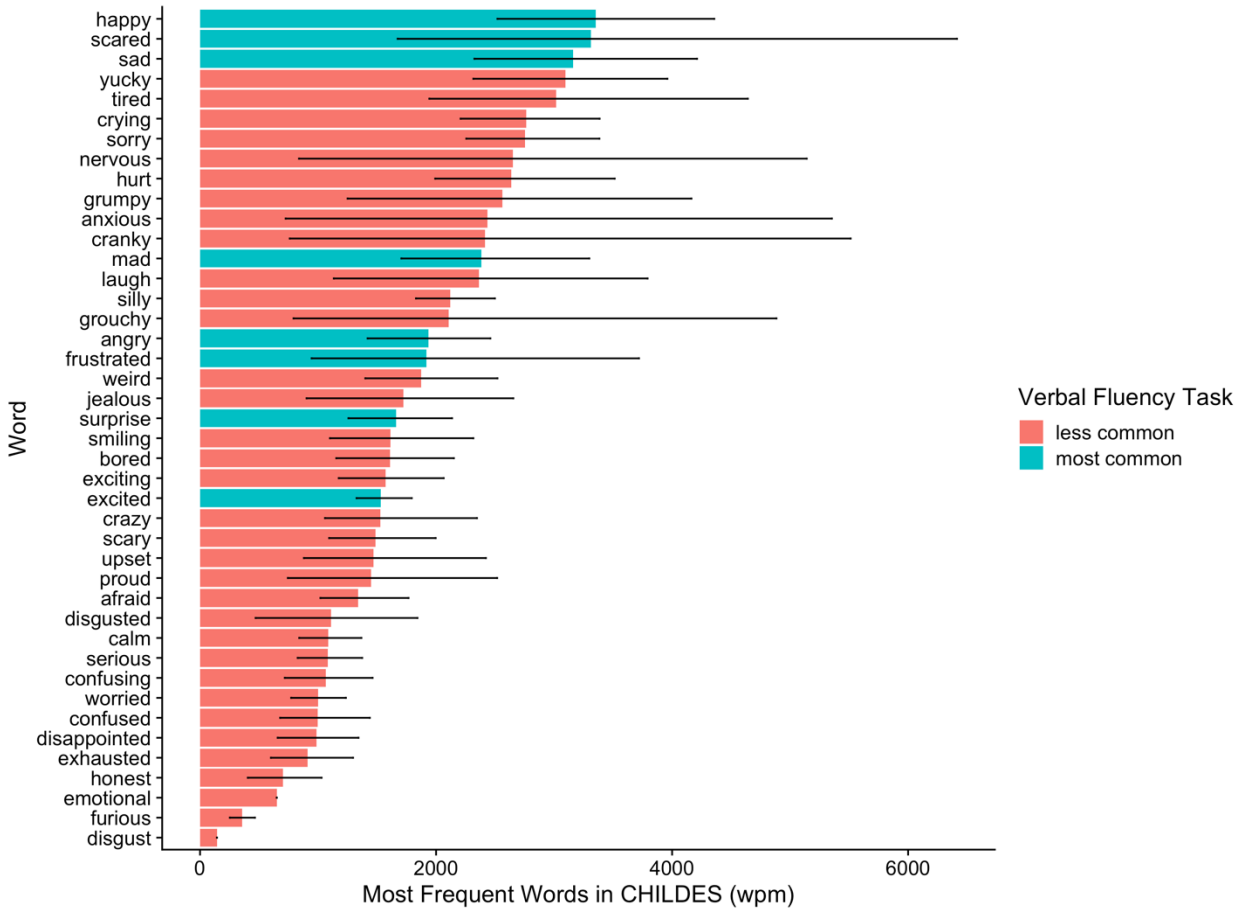
Comparison to CHILDES

Next, I examined whether the most common animal and emotion words produced by children occurred more frequently in child directed speech. To do so, I used the English – North American collection of the CHILDES corpus (MacWhinney, 2000). Many of the words produced by children also frequently appeared in speech to children (see Figures 12-13). Furthermore, words that occurred more frequently in CHILDES (e.g., higher words per million / wpm) were more likely to be produced by children in the emotion ($b = 0.01$, $SE = 0.003$, $t = 2.54$, $p < .02$) and animal ($b = 0.001$, $SE = 0.0003$, $t = 2.05$, $p < .05$) verbal fluency task. The difference in the

amount of animal versus emotion words produced in the task was also reflected in CHILDES, as animal words occurred more frequently than emotion words ($wpm_{emotion}$= 931.13; $wpm_{animals}$=1630), both by frequency (two-sample t-test: $t(10804) = 9.85$, $p<.001$) and raw count (two-sample t-test: $t(12493) = 6.69$, $p<.001$).

Figure 12

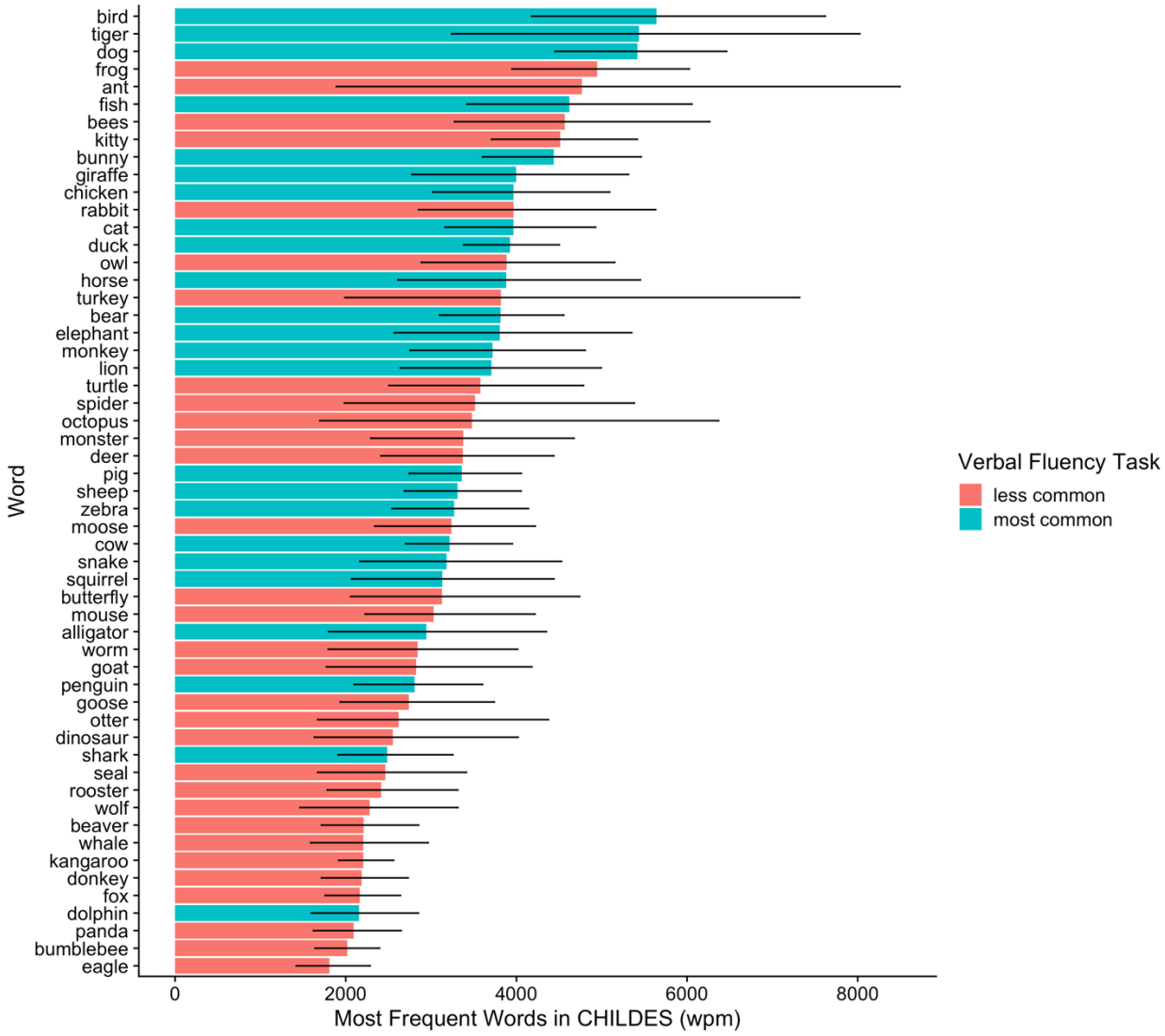
Frequency of Emotion Words per Million in the CHILDES Corpus



Note. Mean frequencies were pulled from all speakers in the English – North American collection. Error bars are 95% confidence intervals computed by nonparametric bootstrapping. Words that were produced by greater than 10% of participants in the verbal fluency task are highlighted blue and labeled as most common.

Figure 13

Frequency of Animal Words per Million in the CHILDES Corpus



Note. Mean frequencies were pulled from all speakers in the English – North American collection. Error bars are 95% confidence intervals computed by nonparametric bootstrapping. Words that were produced by greater than 10% of participants in the verbal fluency task are

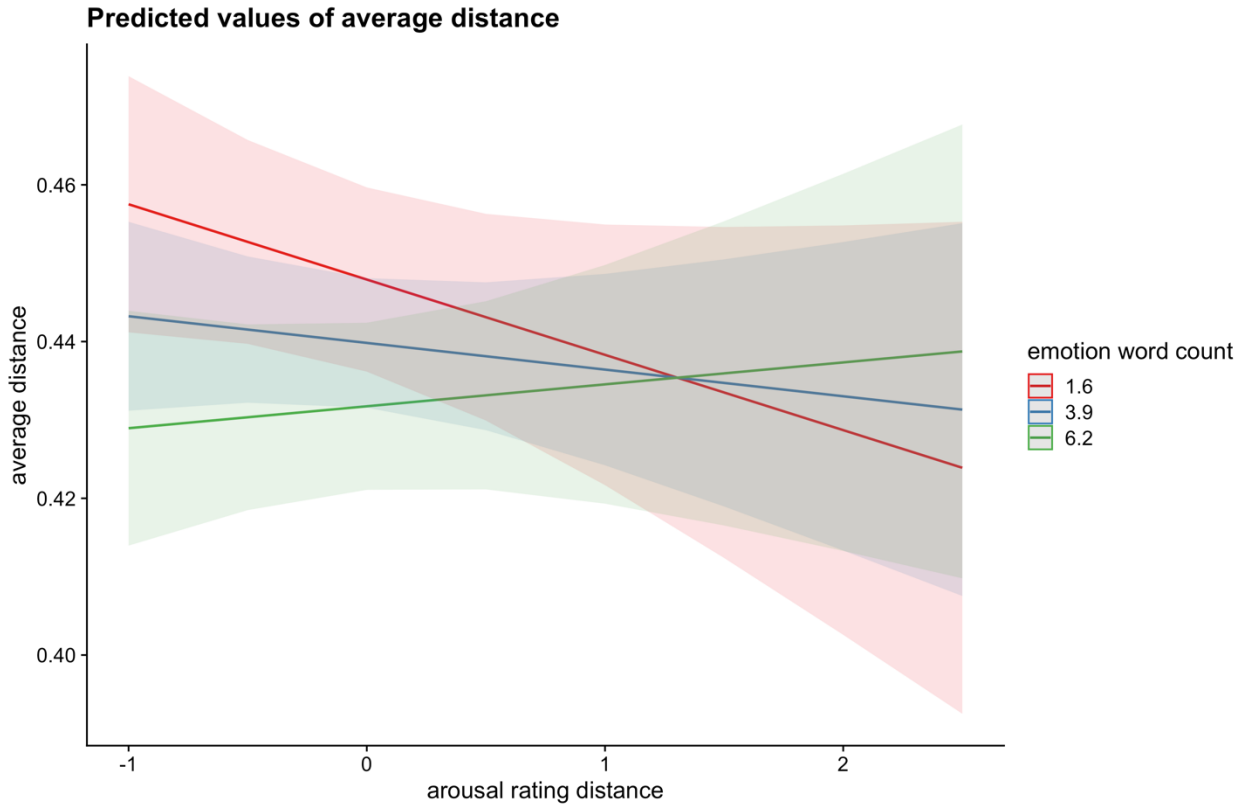
highlighted blue and labeled as most common. For readability, only words that occurred more than 50 times in the corpus are graphed.

Analysis of Verbal Fluency and SpAM

Next, I examined whether the number of emotion words produced in the verbal fluency task related to children's sorting behaviors. First, I examined whether producing more emotion words related to greater use of emotion categories (e.g., happy, sad) and dimensions (e.g., valence, arousal) when sorting. I regressed emotion word count (centered), age (centered), and the dimension of interest (bipolar valence, arousal, positivity, negativity, emotion category match) on the average distance between item pairs (as in the results section: Comparing Dimensions of Affect and Categories in Sorting Behaviors). In all cases, there was no relationship between the number of emotion words produced and using emotion categories or dimensions when sorting (all p's > 0.05). However, there were marginal interactions between emotion word count and arousal (b = 0.003, Wald 95% CI=[0, 0.01], $F(1, 86)= 3.08$, p = .08; Figure 14) and emotion category (b = -0.01, Wald 95% CI=[-0.01, 0.001], $F(1, 86)= 2.70$, p = .10; Figure 15). While the relationships were not significant, trending patterns indicate children who produced more emotion words were more likely to use arousal and emotion category when sorting, even after controlling for age differences.

Figure 14

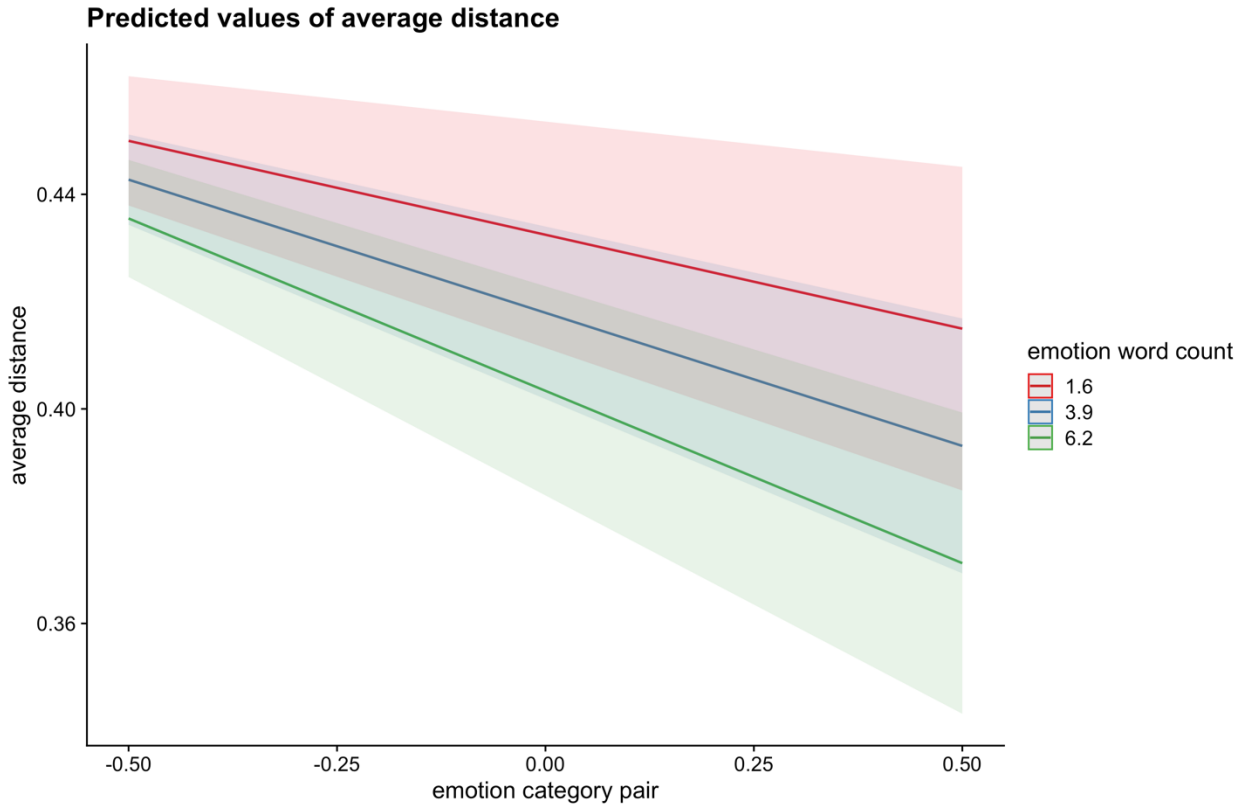
Use of the Arousal Dimension in Sorting Behavior by Children Based on the Number of Emotion Words They Produced



Note. Y-axis represents the average distance between an item pair (e.g., happy-sad). X-axis represents the average arousal rating difference between an item pair (mean centered). Color indicates whether children produced a higher-than-average number of emotion words (green), an average number of emotion words (blue) or a lower-than-average number of emotion words (red). Shading represents 95% confidence intervals. A positively increasing slope indicates that images with similar arousal ratings are placed closer together.

Figure 15

Use of Emotion Categories in Sorting Behavior by Children Based on the Number of Emotion Words They Produced



Notes. Use of emotion categories in sorting behavior by children based on the number of emotion words they produced. Y-axis represents the average distance between an item pair (e.g., happy-sad). X-axis represents whether the pair is from the same emotion category (.5) or different emotion category (-.5). Color indicates whether children produced a higher-than-average number of emotion words (green), an average number of emotion words (blue) or a low number of emotion words (red). Shading represents 95% confidence intervals. A negative slope indicates that images from the same emotion category are placed closer together than images from different emotion categories.

Multidimensional Scaling

Next, I analyzed the verbal fluency task using non-metric multidimensional scaling (NMDS) and hierarchical clustering to try to visualize the semantic structure of the emotion and animal categories. I included all words produced by at least 15% of participants based on previous work examining a verbal fluency task in young children using MDS (Winkler-Rhoades, et al., 2010). The order in which words were produced was used as a measure of distance. To control for differences in the number of words produced (and therefore the maximum "distance"), distance was normalized within each participant. I fit vectors of word ratings for valence and arousal from Warriner et al. (2013), and the percent of children that produced each word, onto the NMDS solution over 1,000 permutations to derive the squared correlation coefficient of each vector (envfit in the R package vegan; Oksanen, 2019). The 2-dimensional solution for emotion had an excellent fit (*Stress 1* < 5%), and the dimensions were marginally related to arousal and valence (Table 10). The 2-dimensional solution for animals had a poorer fit (*Stress 1*<25%) but is comparable to Stress values from previous studies on children's production of animal words (Winkler-Rhoades et al., 2010). The underlying animal dimensions related to arousal, but not valence. For instance, in Figure 16, animals rated as having lower arousal are towards the upper left (e.g., cow, horse) while animals rated as having higher arousal are towards the bottom right (e.g., lion, tiger).

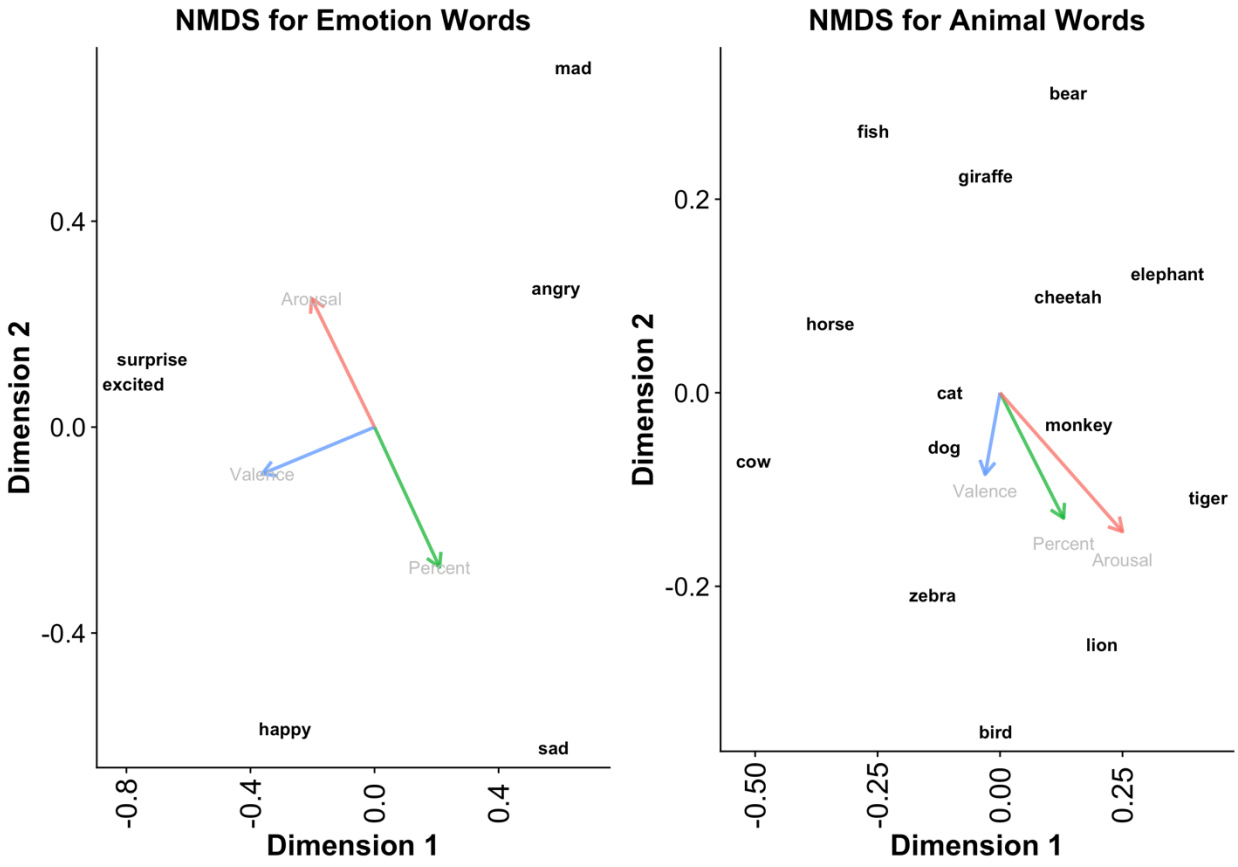
Table 10

Squared Correlation Coefficients for Non-metric MDS Solution

Category	***Dimensions of Interest***		
	Percent Produced	*Valence*	*Arousal*
Emotion Words	0.75	0.89†	0.66†
Animal Words	0.17	0.04	**0.41***

Figure 16

Two-Dimensional Non-metric Multidimensional Scaling Solution of the Emotion and Animal Words Produced by Children in the Verbal Fluency Tasks



Note. Vectors represent the squared correlation coefficient of the underlying dimensions to ratings of valence (blue arrow), arousal (red arrow), and the percent of children that produced each word (green arrow).

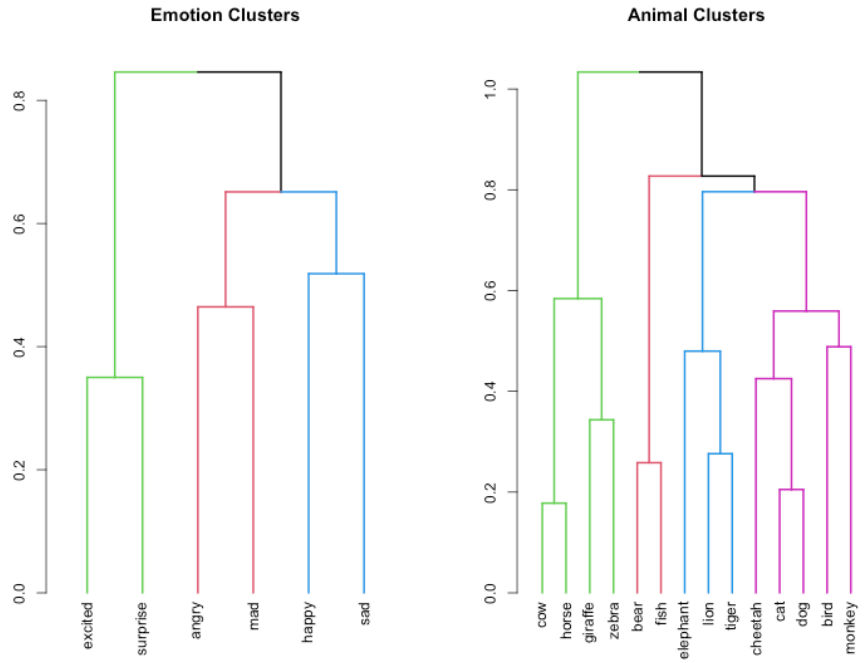
Hierarchical Clustering

Last, I examined which clusters emerged in children's production of emotion and animal words (hierarchical clustering, Ward's Method) using the distance between words produced

during the verbal fluency task (Figure 17). For emotion words, the clusters were based on how frequently children produced the words (F(2,3)=26.44, *p*=.01). The most frequent words (mad, angry, happy, sad) occured together compared to the least frequent words (excited, surprised). These clusters occurred after controling for differences in the amount of words produced by normalizing the distances within each participant. For animals, the clusters did not explain variance in the percent of children that produced each word, valence, or arousal (all p's > 0.19).

Figure 17

Dendrograms of the Emotion and Animal Words Listed by Children in the Verbal Fluency Task



Note. Based on changes in the height of the dendrograms at different clusters, the 3-cluster solution is highlighted for emotion words, and the 4-cluster solution is highlighted for animal words. The emotion word clusters are related to how often children produced the words during the task (blue = most frequent, red = medium frequent, green = least frequent).

General Discussion

This study reveals developmental changes in how humans represent perceptual information associated with facial cues of emotions. By using a non-verbal, open-ended procedure, I circumvented a number of traditional limitations incurred in the assessment of knowledge of facial cues in young children. I found that children and adults primarily rely upon the affective dimension of valence. Adult-like reliance on common English language emotion categories (happy, sad, angry, etc.) emerged only gradually, with little evidence that children consistently used these categories until around five years of age. Similar patterns of incremental change in how children represent facial cues of emotion emerged in both supervised and unsupervised analyses.

Nuances in the Use of Valence

Valence accounted for a very large proportion of all participant's sorting behaviors, providing converging evidence with prior studies (e.g., Jackson et al., 2019; Nook et al., 2017). Negativity even explained 3-year-olds' sorting behaviors, although this result should be interpreted cautiously given how this age group approached the practice sort. My findings of an early role for negativity in children's emerging emotion knowledge is consistent with reports that young children display greater knowledge of negative emotions (Lagattuta & Wellman, 2001), attend more to negative faces (Lagattuta & Kramer, 2017), engage in greater discussion of negative emotions (Lagattuta & Wellman, 2002), and voluntarily explore negatively-valenced stimuli (Grisanzio et al., 2020). Similarly, children produced more negative words in the emotion verbal fluency task in the present study.

The present data also reveal new insights about valence. First, treating valence as bivariate (represented by separate unipolar scales of positivity and negativity) better accounted

for sorting behavior than treating it as a bipolar construct (a single continuum ranging from positive to negative), converging with findings surrounding the structure of the affect system (e.g., Cacioppo et al, 1999; Norris, et al., 2010). Second, positivity and negativity are not used equally early in development. Young children relied heavily on negativity, did not consistently use positivity, and saw greater increases in the explanatory power of the models when bivariate valence was used in place of bipolar valence. This difference is not explained entirely by the stimuli used in the study, as there were relatively even numbers of images with more negative (anger, fear, disgust, neutrality) and positive (happy, calm, excited, surprised) ratings. Last, allowing positivity and negativity to exist separately might also better capture human experience: One can experience spicy food as both painful and delicious, or horror movies as both frightening and entertaining (although see Russell, 2017 for a critique on how bivariate valence may play a role in judgments about affect but not experienced affect).

The Limits of Arousal

Though often discussed in tandem with valence, I found that arousal was only sometimes related to children's sorting behaviors, explained a much smaller proportion of behavior than valence, and did not consistently correlate with any multidimensional scaling solution. This limited role of arousal is in contrast with many theories of emotion that posit that emotions initially emerge from a 2-factor understanding of valence and arousal (for reviews see Barrett & Bliss-Moreau, 2009; Russell & Barrett, 1999). There are a number of reasons for these divergent conclusions. First, arousal can be presented to research participants in different ways—such as perceptions of excitement, activation, or intensity in the self or others—that elicit varying interpretations. It is also sometimes not even measured at all. For instance, observing multidimensional scaling solutions of emotion words led some researchers to draw conclusions

about the “immutability of valence and arousal in the foundation of emotion” (Bliss-Moreau et al., 2020; Nook et al., 2017) without ever measuring arousal with any kind of ratings.

Second, arousal may be better captured as a measure of individual difference. There is a large amount of individual variation in the reporting of arousal (Barret & Bliss-Moreau, 2009; Kuppens et al., 2013), and self—but not experimenter—ratings of arousal predict attentional differences to emotional stimuli (Sutherland & Mather, 2018). Furthermore, perceptions of emotional intensity appear to be highly variable, and shift depending on the context and statistical information encountered (Plate, Wood, et al., 2019; Woodard, Plate, & Pollak, 2021), and in response to training (Leitzke, et al., 2020). Taken together, this body of work suggests that arousal may be less of a static trait underlying emotion perception, and rather one that is sensitive to context and individual experience.

Third, arousal may index natural covariation in positivity and negativity, rather than capturing unique variance in emotion (Haj-Ali et al., 2020; Kron et al., 2013). The data provide some support for this possibility, as arousal is no longer a significant predictor of sorting behavior among 5-year-olds, 6-year-olds, and adults when bivariate valence is included. Furthermore, positivity and negativity (rather than valence and arousal) better capture the underlying dimensions of the multidimensional scaling solutions in the present study. This reframing of arousal as covariation in positivity and negativity pairs well with current views on the autonomic nervous system. Rather than generalized states of arousal (from sympathetic activity) and non-arousal (from parasympathetic activity), activation and deactivation are separable and can occur together or separately to varying degrees (Berntson, et al., 2008; Norman, et al., 2014).

Verbal fluency task

The verbal fluency task suffered from data loss—only 75 of 107 participants produced any emotion words at all. For children who did participate, the task was often quite difficult. One reason for this data loss and difficulty could be that the ability to switch between categories and subcategories during the task (e.g., switch from saying farm animals to zoo animals) is linked to executive function ability (Amunts, et al., 2020; Hirshorn & Thompson-Schill, 2006; Snyder & Munakata, 2010), which is still developing in young children (Huttenlocher & Dabholkar, 1997). Thus, the task may not only be testing semantic knowledge – but issues in accessing this knowledge tied to executive function development. Related to this possibility, younger children in the present study were more likely to not participate, and older children were more likely to produce more animal and emotion words.

Despite some difficulties relating to data loss and task difficulty, many interesting patterns emerged. The words produced by children tended to be those that occurred more frequently in child directed speech—including more animal than emotion words. The greater frequency of animal words in both child speech and the verbal fluency task could be because animals are more concrete categories with more reliable perceptual cues, while many emotion categories are more abstract with less reliable perceptual cues (Hoemann et al., 2020), and "fuzzy" categorical boundaries (Cowen & Keltner, 2017).

While there were some marginal relationships regarding arousal, valence, and emotion categories, there was not strong evidence that the emotion words produced predicted differences in sorting behavior, or that the underlying dimensions (from MDS) were the same across the verbal fluency and SpAM tasks. Similarly, clusters derived from emotion word production were unrelated to dimensions of affect. These different findings across tasks could be related to task

difficulty and data loss, as well as differences in the processing of emotion pictures versus emotion words. Neuroimaging studies on emotional pictures versus emotion words have found that the different stimuli elicit different brain activation patterns (Feng et al., 2021). Future work could further examine this possibility by comparing whether using SpAM with emotion words has a stronger relationship with the verbal fluency task.

Last, the consistency with which children produced certain emotion words was notable. Nearly all children who completed the task stated “happy”, “sad” and “mad/angry”. By comparison, the most frequent animal category (“lion”) was produced by only half of participants. The consistency with which the words were produced raises the possibility that perhaps these early terms are semantic “seed words” that lay the groundwork for later emotion vocabulary development (e.g., Babineau, et al., 2021; Christophe, et al., 2016). This consistency appears across studies, as children produce certain emotion words earlier and more accurately (e.g., happy, sadness, and anger) than others (e.g., surprise and disgust; Widen, 2013).

What Changes in the Representation of Facial Cues of Emotion?

The present study captured changes in the use of information that underlies children’s sorting behaviors; however, it’s difficult to pinpoint what is driving these changes. Children at this age are gaining more life experience generally, beginning school, and continuing to hone their language and reading skills. Similarly, neural developments are leading to improvements in working memory, attention, and cognitive control. The present section focuses on how perceptual and linguistic experience are likely impacting how children approach the practice and facial sorts across development.

Task Understanding

Across development, children's general task understanding improved. 3-year-olds during the practice phase did not appear to use superordinate categories to guide their similarity judgments. Rather, children in this age range appeared to approach the task with many different strategies: some did appear to use superordinate categories like animals and vehicles (N = 5), some separated each image far apart, viewing each stimulus as distinct (N = 4), some appeared to focus on other perceptual features, like color, to guide their decisions (N = 5), and some appeared to sort at random (N = 7). However, observations in this regard are subjective and future research will need to adjust the task to better understand and tease apart these possibilities. In future studies, it would be interesting to examine how 3-year-olds approach simpler sets of stimuli, such as shapes varying in size and color. This would allow researchers a better understanding of how this age group understands the task, and which features guide their approach. Last, from a practical standpoint, data collection for 3-year-olds was most impacted by the COVID-19 outbreak, as this age bin had the smallest sample size (N=21), which reduced statistical power for this age group.

The Role of Dimensions of Affect

Another change that occurred across development was the manner in which different dimensions of emotion explained children's sorting behaviors. Categories represented by multiple underlying dimensions take longer to learn (Sanborn et al., 2021), and children's changes in their use of dimensions over time may reflect this difficult learning problem. Changes in which dimensions explained children's organization of facial cues did not seem to reflect children's responses becoming more consistent or general task competency. With the exception of the 3-year-old age group, children demonstrated good comprehension of the task during the

practice phase and sorted items unrelated to emotion similarly to adults. Rather, beginning at 4 years of age, children systematically organized facial configurations according to broader dimensions, with some dimensions (e.g., valence) gaining increasing explanatory weight and other, initially influential dimensions (e.g., arousal) diminishing in effect size among children. If children were simply becoming more similar to adults, then one would expect to see a lot of noise in child behavior that reduces bit by bit until adult-like behavior emerges. However, we found that how children prioritized and used emotional information was distinctive: children closer in age had clustering structures that were much more aligned with one another than with those of adults. These results suggest that children prioritized perceptual information about emotion in a systematic manner that was distinct from how adults organized this same information.

The Role of Emotion Categories

Another change was the emergence of the use of emotion categories to guide sorting behavior. Common English-language emotion categories guided the sorting behaviors of adults and children beginning around age 5. This change in the use of emotion category could reflect that children first use broad, primarily valence-based distinctions, and with greater experience, draw more fine-grained distinctions that use emotion category information (Matthews et al., 2020; Widen, 2013). This pattern is similar to those discovered in other domains of development. For instance, the development of non-emotional categories (e.g., animals and other natural kinds) reveals that children first make broad distinctions (e.g., birds vs. mammals) and later show finer differentiation of items based on their category membership (e.g., flamingoes vs. peacocks; Vales, Stevens, et al., 2020). However, these findings contradict some infant research, which finds that discrete emotion categories emerge earlier than superordinate categories like valence

(Ruba et al., 2017, 2020; White et al., 2019). This discrepancy may be due to methodological differences, as infant research focuses more on perceptual discrimination (for a full discussion of this issue, see Ruba & Pollak, 2020), rather than graded similarity judgments.

Alternatively, this apparent contradiction could be an artifact of study designs. Valence and discrete categories are often pitted against one another, but the present study allowed children to use both at the same time without having the two sources of information compete. This design found that valence and emotion categories are often related—for instance, anger and disgust had the most negative valence ratings, while happy faces tended to have some of the most positive valence ratings. Thus, knowledge of valence can often give a learner traction on knowledge that appears to be category-related, and vice versa. The increased use of both valence and category information across development supports this reciprocal relationship. In fact, rather than a distinct shift from using valence to using emotion categories, the study found continued and refined use of both across development. In this way, I do not think that the evidence of emotion development is "broad to narrow" or "narrow to broad", rather it's both the broad (e.g., affective dimensions like valence) and the narrow (e.g., emotion categories) guiding one another. It's possible that particularly salient perceptual (e.g., highly emotional faces) and linguistic stimuli (e.g., seed words) may guide learning about both the individual categories and the broader dimensions simultaneously.

Children's Conceptual Development

The changes observed in children's behavior could also reflect transitions in conceptual development. Children may shift from more perceptual, similarity-based categories to categories shaped more by rules and labels. These labels in turn, promote the development of conceptual networks and hierarchies that do not seem to appear until later in development (for a review see

Sloutsky & Deng, 2019). For instance, the present study found the use of emotion categories emerged around 5 years of age, and that children's dendrograms gradually became more aligned with adults' over time. This shift in conceptual development is reflected in the growth of children's emotion vocabulary around this time, as younger children (i.e., 4-year-olds) initially rely more on example situations when discussing emotion words and shift over time to provide more general definitions, abstract explanations, and synonyms (Nook et al., 2020).

Perceptual Learning from Distributions

Another possible source of development is the degree to which perceptual information is guiding children's understanding of facial cues of emotion. Children may use statistical regularities in their perceptual environment to guide their socioemotional learning (for a review see Plate, et al., 2021). As children encounter a greater variety of individuals at school and in other areas, their recognition of facial cues may also improve—although there is some debate as to whether perceptual cues of emotion are reliable signals at all (e.g., Barrett et al., 2019; Le Mau et al., 2021). Emotions seem to be "fuzzier", more probabilistic categories (Cowen & Keltner, 2021) in which perceptual cues signal greater than chance information, without being perfectly informative.

Similar to speech perception, facial and vocal signals of emotion may suffer from a "lack of invariance" (Liberman, 1957; Liberman et al., 1967), which is to say that there do not seem to be reliable mappings between phonemes and different acoustic cues (in the case of speech), or between emotions and different facial and vocal cues (in the case of emotion). A phoneme can sound very different depending on who is saying it, the context, speech conditions, and random errors in speech production; yet individuals are highly accurate at perceiving these sounds. One of the mechanisms that supports this perception is rapid adaptation via tracking the distributional

properties of vocal input (Kleinschmidt & Jaeger, 2011, 2015; Samuel & Kraljic, 2009). This rapid adaptation is not simply an in the moment shift in signal - it appears to update perceivers' underlying representations (Clarke-Davidson, et al., 2008; Dahan, et al., 2008), and over time allows perceivers to use these variations in speech input to shift categorization processes when necessary (Weatherholtz & Jaeger, 2016). Similar rapid adaptation occurs when adjusting to the facial and vocal expressivity of different actors (Plate, Wood, et al, 2019; Woodard, Plate, Morningstar, et al, 2021; Woodard, Plate & Pollak, 2021).

Similarities between "talker" and "expresser" variability present an opportunity for emotion to utilize advances in our understanding of speech perception to uncover more about children's improvements in emotion recognition. For instance, models of speech perception suggest that much of the variability in speech perception is not random – rather the variability is predicted by many social variables like age, gender, and dialect (Kleinschmidt, 2019). Similarly, variation in facial cues of emotion may be reliably predicted by social variables, and children may use this information to guide their perceptual learning of emotion categories.

Implications for Theories of Emotion

Several explanations commonly used to account for the emergence of emotion find only partial support in the present data. Young children in the task did not begin to use basic emotion categories until around the age of 5, arguing against the theory that this knowledge plays a large role in young children's emotion understanding (Izard, 2007). Children also did not rely equally on the dimensions of valence and arousal, instead using negative valence far more heavily, lending less support to the model of Core Affect (Russell & Barrett, 1999). Children's emotion word production did not predict differences in sorting by emotion categories or dimensions, as might be predicted by constructivists who argue for a strong role of language in emotion

understanding (e.g., Hoemann et al., 2019). The picture of emotion development that emerges from the data is of an incremental learning process in which children change their representations of emotion using combinations of factors that are weighted differently across development. A focus on how children learn about emotion categories could help to guide theories towards a greater understanding of the nature of emotion.

Limitations

There are a number of limitations to the present work, particularly with regard to the set of stimuli used. The facial stimuli used were typical for this area of research and were selected in part to make the present study easier to compare to other studies. There was some degree of variation and diversity in the stimuli used, as there were open and closed mouthed images of nine emotion categories, from Asian, Black, and White males and females. However, to have greater generalizability and a fuller understanding of emotional development, future studies require (a) more variety in the age, gender, ethnic, and racial identities of the individuals providing emotion cues; (b) use of emotion categories beyond those commonly used in English; (c) stimuli that are naturally occurring rather than posed; and (d) less reliance on faces alone and more emphasis on the situational contexts and broader variety of dynamic visual and auditory stimuli that characterize human interactions (Srinivasan & Martinez, 2018; Woodard, Plate, Morningstar, et al., 2021).

While a greater variety of stimuli would strengthen claims of generalizability in the present task, there are reasons to be optimistic. For instance, there were similar patterns of behavior across both the Same and Different Individual sorts, and gender did not guide participants' sorting behavior in the Different Individual sort. Similarly, using young adult faces rather than children's faces does not necessarily make children less able to interpret facial cues.

In the area of emotional prosody, youth listeners are more accurate at identifying the emotions of adult speakers than they are at identifying the emotions of other youth speakers, and adult listeners are better than youth listeners at identifying the emotions of both youth and adult speakers (Morningstar, et al., 2018), suggesting that adults tend to display clearer emotional signals overall.

Other limitations of the present task were decisions regarding experimental design and analysis. The operationalization of emotion category in this study may have underrepresented its influence on sorting behaviors. The current study design was selected based on how emotion categories are traditionally represented in the study of children's emotional development—that is emotion categories were assumed to have clear boundaries and to represent only one emotion at once. However, there is some evidence that emotion categories are "fuzzier" and that when faces represent multiple emotion categories at once, they may do a better job of representing knowledge about facial cues of emotion than valence and arousal (Cowen & Keltner, 2017).

Furthermore, the verbal fluency task—as a measure of semantic knowledge—did not appear to be appropriate for younger children given participants' stress and the amount of data loss. Additionally, distance in the verbal fluency task was measured as the difference in the order at which words were said, rather than reaction time, which is often a more sensitive and preferred measurement. Order was used because the task was made more interactive to alleviate participant stress, making reaction time less reliable. Using the verbal fluency task with older children and with reaction time as a measure of distance would likely be a more informative indicator of children's semantic knowledge.

Future Directions

Future research should explore how children construe experimental tasks such as the one used here. One advantage of the current task is that children are given minimal verbal prompting to guide their sorting behavior, allowing the study of children's spontaneous emotion judgments. However, research on conceptual development reveals that even subtle variations in task context, such as the verbal prompts used to introduce the task, can reveal different facets of children's knowledge (e.g., Deák & Bauer, 1995; Gentner & Christie, 2014; Waxman & Namy, 1997). For instance, future studies could ask children to focus on specific features ("Focus on people's eyes to figure out what they are feeling"), frame the task using specific emotion labels ("Think about if they're feeling happy or mad"), or present children with different perceptual information before the task (e.g., manipulating children's experience with the facial expressivity of different actors as in Woodard, Plate, & Pollak, 2021). Together, these variations would help us to better understand the different conceptual and perceptual influences with which children approach the task. Similarly, children could sort emotion cues from a variety of modalities (facial cues, vocal cues, and emotion words). The consistency in use of information across all of these framings would lend strong support to task-independent representations of perceptual information about emotion, while variability would provide evidence for context-sensitive use of different factors when children evaluate emotion cues.

In addition to differences in framing, the task could also be used as a more sensitive measure to assess longer term differences in emotion experience. For instance, given the role that fiction might play in emotion learning (Schwering et al., 2021), the SpAM could be used to assess emotion knowledge before and after a multiweek exposure to fiction with more emotive

contexts compared to fiction with less emotive contexts (as in Unger & Fisher, 2019). Together, these lines of work will help researchers to better understand how children learn about and organize perceptual cues of emotion, and how they balance flexibility and stability in this process.

Last, future studies could focus on better understanding individual differences in how children approach the task. Similarities between children and adults on the practice sort versus the facial sorts suggest that the task is accessing unique information about emotion development rather than improvement on the task generally. A focus on individual differences in the task could uncover differences in how individuals use positivity and negativity when judging emotional cues traditionally considered more ambiguous. There are stable individual differences in whether certain ambiguous expressions are more likely to be labeled as positive or negative (Neta, et al., 2017, 2018), some of which may be reflective of differences in family expressivity, anger exposure (Plate, Bloomberg, et al., 2019), and abuse (Pollak et al., 2000). In this way, SpAM could provide a better understanding of how differences in past experience may come to alter emotion recognition.

Conclusion

Emotions are critical for human adaptation and survival, yet relatively little is understood about how humans come to understand and represent emotion concepts. SpAM and the verbal fluency task both present fruitful avenues for more diverse ways to study emotion perception and representation. Graded similarity judgments from SpAM allow researchers to examine relationships across emotion stimuli, to examine which dimensions and categories might underlie emotion understanding, and to address these questions using a variety of statistical methods (regression, clustering, and multidimensional scaling). Findings from this novel variation of

SpAM were largely convergent with the existing literature—children and adults use valence, arousal, and emotion category to guide their judgments of facial cues—while raising nuanced points. Valence and arousal do not equally influence emotion judgments, and emotion categories do not guide behavior until later in development.

The present study also touches on critical debates in the field of emotion. With debates raging on what – if anything – can be learned from perceptual cues of emotion, the present task demonstrates that across adults and children there is convergence in the information extracted from these cues. SpAM is able to guide this research with fewer assumptions of the essential "truth" of an experimenter assigned label. The picture of emotion development that emerges from the data is of an incremental learning process in which children change their representations of emotion using combinations of factors that are weighted differently across development. This insight opens the door for new investigations about how humans learn to navigate the complex communicative system of the social world.

References

Ameel, E., Malt, B., & Storms, G. (2008). Object naming and later lexical development: From baby bottle to beer bottle. *Journal of Memory and Language*, *58*(2), 262-285.

Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S., & Weis, S. (2020). Executive functions predict verbal fluency scores in healthy participants. *Scientific reports*, *10*(1), 1-11. https://doi.org/10.1038/s41598-020-65525-9.

Babineau, M., de Carvalho, A., Trueswell, J., & Christophe, A. (2021). Familiar words can serve as a semantic seed for syntactic bootstrapping. *Developmental Science*, *24*(1), e13010.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in psychology*, *8*, 456.

Bakdash, J.Z., & Marusich, L. R. (2021). rmcorr: Repeated Measures Correlation. R package version 0.4.3. https://CRAN.R-project.org/package=rmcorr

Baker, F. B. (1974). Stability of two hierarchical grouping techniques case I: sensitivity to data errors. *Journal of the American Statistical Association*, *69*(346), 440-445.

Bates, D., Mächler, M., Bolker, B., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baron-Cohen, S., Golan, O., Wheelwright, S., Granader, Y., & Hill, J. (2010). Emotion Word Comprehension from 4 to 16 Years Old: A Developmental Survey. *Frontiers in Evolutionary Neuroscience*, *2*, 25. https://doi.org/10.3389/fnevo.2010.00109

Barrett, L. F., & Bliss-Moreau, E. (2009). Chapter 4: Affect as a Psychological Primitive. In *Advances in Experimental Social Psychology* (Vol. 41, pp. 167–218). Academic Press. https://doi.org/10.1016/S0065-2601(08)00404-8

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397. https://doi.org/10.1016/j.cognition.2013.02.011

Berntson, G. G., Norman, G. J., Hawkley, L. C., & Cacioppo, J. T. (2008). Cardiac autonomic balance versus cardiac regulatory capacity. *Psychophysiology*, *45*(4), 643-652.

Bliss-Moreau, E., Williams, L. A., & Santistevan, A. C. (2020). The immutability of valence and arousal in the foundation of emotion. *Emotion*, *20*(6), 993.

Boehmke, B., & Greenwell, B. (2020). *Hands-on machine learning with R*. Chapman and Hall, CRC Press.

Braginsky, M., Sanchez, A., & Yurovsky, D. (2020). childesr: Accessing the 'CHILDES' Database. R package version 0.2.1. https://CRAN.R-project.org/package=childesr

Bridges, K. M. B. (1932). Emotional Development in Early Infancy. *Child Development*, *3*(4), 324-341. https://doi.org/10.2307/1125359

Brody, L. R. (2000). The socialization of gender differences in emotional expression: Display rules, infant temperament, and differentiation. *Gender and emotion: Social psychological perspectives* (pp. 24-47). Cambridge University Press, New York, NY.

Brody, L. R., & Hall, J. A. (2010). Gender, emotion, and socialization. In *Handbook of gender*

research in psychology (pp. 429-454). Springer, New York, NY.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of personality and Social Psychology*, *76*(5), 839.

Caron, R. F., Caron, A. J., & Myers, R. S. (1985). Do Infants See Emotional Expressions in Static Faces? *Child Development*, *56*(6), 1552-1560. https://doi.org/10.2307/1130474

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive science*, *38*(2), 383-397.

Christophe, A., Dautriche, I., de Carvalho, A., & Brusini, P. (2016). Bootstrapping the syntactic bootstrapper. In *Proceedings of the 40th Annual Boston University Conference on Language Development* (pp. 75-88). Somerville, MA: Cascadilla Press.

Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., & Sonuga-Barke, E. J. (2015). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *British Journal of Developmental Psychology*, *33*(2), 218-236.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception and Psychophysics*, *70*(4), 604–618. https://doi.org/10.3758/PP.70.4.604

Coburn, A., Kardan, O., Kotabe, H., Steinberg, J., Hout, M. C., Robbins, A., MacDonald, J., Hayn-Leichsenring, G., & Berman, M. G. (2019). Psychological responses to natural patterns in architecture. *Journal of Environmental Psychology*, *62*, 133-145.

Cole, P. M., Armstrong, L. M., & Pemberton, C. K. (2010). The role of language in the development of emotion regulation. In S. D. Calkins & M. A. Bell (Eds.), *Child*

development at the intersection of emotion and cognition (pp. 59–77). American Psychological Association. https://doi.org/10.1037/12059-004

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75-93.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38).

Cowen, A. S., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, *25*, 124-136.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*(3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2020). Statistical power for cluster analysis. *ArXiv*. http://arxiv.org/abs/2003.00381

Deák, G., & Bauer, P. J. (1995). The effects of task comprehension on preschoolers′ and adults′ categorization choices. *Journal of Experimental Child Psychology*, *60*(3), 393-427.

De Vries, A. & Ripley, B. D. (2020). ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'. R package version 0.1.21. https://CRAN.R-project.org/package=ggdendro

Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., Clément, F., Cunningham, W. A., Damasio, A., Damasio, H., D'Arms, J., Davidson, J.W., de Gelder, B., Deonna, J., de Sousa, R., … & Sander, D. (2021). The rise of affectivism. *Nature Human Behavior,* 5, 816–820 (2021).

https://doi.org/10.1038/s41562-021-01130-8

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101-107.

Feng, C., Gu, R., Li, T., Wang, L., Zhang, Z., Luo, W., & Eickhoff, S. B. (2021). Separate neural networks of implicit emotional processing between pictures and words: A coordinate-based meta-analysis of brain imaging studies. *Neuroscience & Biobehavioral Reviews*.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, *78*(383), 553-569.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718-3720. https://doi.org/10.1093/bioinformatics/btv428

Gobbo, C., & Chi, M. (1986). How knowledge is structured and used by expert and novice children. *Cognitive development*, *1*(3), 221-237.

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

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125-157.

Grisanzio, K. A., Sasse, S. F., Nook, E. C., Lambet, H. K., McLaughlin, K. A., & Somerville, L. H. (2021). Voluntary pursuit of negatively valenced stimuli from childhood to early adulthood. *Developmental Science*, *24*(2), e13012. https://doi.org/10.1111/desc.13012

Haj-Ali, H., Anderson, A. K., & Kron, A. (2020). Comparing three models of arousal in the human brain. *Social Cognitive and Affective Neuroscience*, *15*(1), 1–11. https://doi.org/10.1093/scan/nsaa012

Hall, J. A., & Gunnery, S. D. (2013). 21 Gender differences in nonverbal communication. In *Nonverbal Communication* (pp. 639-670). Berlin: Mouton de Gruyter.

Harré, R. (Ed.). (1986). *The social construction of emotions*. Blackwell.

Hirshorn, E. A., & Thompson-Schill, S. L. (2006). Role of the left inferior frontal gyrus in covert word retrieval: neural correlates of switching during verbal fluency. *Neuropsychologia*, *44*(12), 2547-2557.

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology*, *55*(9), 1830–1849. https://doi.org/10.1037/dev0000686

Hoemann, K., Wu, R., LoBue, V., Oakes, L. M., Xu, F., & Barrett, L. F. (2020). Developing an understanding of emotion categories: Lessons from objects. *Trends in Cognitive Sciences*, *24*(1), 39-51.

Hoemann, K., Vicaria, I. M., Gendron, M., & Stanley, J. T. (2021). Introducing a face sort paradigm to evaluate age differences in emotion perception. *The Journals of Gerontology: Series B*, *76*(7), 1272-1281.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*(1), 256–281. https://doi.org/10.1037/a0028860

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93-103.

Hout, M. C., Godwin, H. J., Fitzsimmons, G., Robbins, A., Menneer, T., & Goldinger, S. D. (2016). Using multidimensional scaling to quantify similarity in visual search and beyond. *Attention, Perception, and Psychophysics*, *78*(1), 3–20. https://doi.org/10.3758/s13414-015-

1010-6

Hout, M. C., & Goldinger, S. D. (2016). SpAM is convenient but also satisfying: Reply to verheyen et al. (2016). *Journal of Experimental Psychology: General*, *145*(3), 383–387. https://doi.org/10.1037/xge0000144

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193-218.

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *Journal of comparative Neurology*, *387*(2), 167-178.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, *2*(3), 260-280.

Jackson, J. C., Watts, J., Henry, T. R., List, J. M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2017). Why are faces denser in the visual experiences of younger than older infants? *Developmental psychology*, *53*(1), 38.

Jayaraman, S., & Smith, L. B. (2019). Faces in early visual environments are persistent not just frequent. *Vision research*, *157*, 213-221.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Keltner, D., Tracy, J. L., Sauter, D., & Cowen, A. (2019). What Basic Emotion Theory Really Says for the Twenty-First Century Study of Emotion. *Journal of Nonverbal Behavior*, *43*(2), 195–201). https://doi.org/10.1007/s10919-019-00298-y

Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic

recalibration and selective adaptation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10-19)

Kleinschmidt, D. F., & Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698

Koch, A., Speckmann, F., & Unkelbach, C. (2020). Q-SpAM: How to efficiently measure similarity in online research. *Sociological Methods & Research*, 0049124120914937. https://doi.org/10.1177/0049124120914937

Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: expression, experience, and physiology. *Journal of Personality and Social Psychology*, *74*(3), 686-703.

Kron, A., Goldstein, A., Lee, D. H. J., Gardhouse, K., & Anderson, A. K. (2013). How Are You Feeling? Revisiting the Quantification of Emotional Qualia. *Psychological Science*, *24*(8), 1503–1511. https://doi.org/10.1177/0956797613475456

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, 1-27.

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological bulletin*, *139*(4), 917.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(1), 1-26.

LaBounty, J., Wellman, H. M., Olson, S., Lagattuta, K., & Liu, D. (2008). Mothers' and fathers'

use of internal state talk with their young children. *Social Development*, *17*(4), 757-775.

Lagattuta, K. H., & Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Development*, *72*(1), 82 102.

Lagattuta, K. H., & Wellman, H. M. (2002). Differences in early parent-child conversations about negative versus positive emotions: implications for the development of psychological understanding. *Developmental Psychology*, *38*(4), 564.

Lagattuta, K. H., & Kramer, H. J. (2017). Try to look on the bright side: Children and adults can (sometimes) override their tendency to prioritize negative faces. *Journal of Experimental Psychology: General*, *146*(1), 89.

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 & Emotion*, *23*(3), 453–480. https://doi.org/10.1080/02699930801994054

Larsen, J. T., & McGraw, A. P. (2011). Further evidence for mixed emotions. *Journal of Personality and Social Psychology*, *100*(6), 1095–1110. https://doi.org/10.1037/a0021846

Laukka, P., & Elfenbein, H. A. (2021). Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review*, *13*(1), 3-11.

Leitzke, B. T., & Pollak, S. D. (2016). Developmental changes in the primacy of facial cues for emotion recognition. *Developmental Psychology*, *52*(4), 572.

Leitzke, B. T., Plate, R. C., & Pollak, S. D. (2020). Training reduces error in rating the intensity of emotions. *Emotion*. Advance online publication. https://doi.org/10.1037/emo0000763

Le Mau, T., Hoemann, K., Lyons, S. H., Fugate, J., Brown, E. N., Gendron, M., & Barrett, L. F. (2021). Professional actors demonstrate variability, not stereotypical expressions, when

portraying emotional states in photographs. *Nature communications*, *12*(1), 1-13.

Leppänen, J. M., & Nelson, C. A. (2009). Tuning the developing brain to social signals of emotions. *Nature Reviews Neuroscience*, *10*(1), 37–47). NIH Public Access. https://doi.org/10.1038/nrn2554

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America*, *29*(1), 117-123.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Lindquist, K. A. (2021). Language and Emotion: Introduction to the Special Issue. *Affective Science*, 1-8.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*, *66*(3), 516-553.

Lutz, C., & White, G. M. (1986). The anthropology of emotions. *Annual Review of Anthropology*, *15*(1), 405-436.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates

Mattek, A. M., Burr, D. A., Shin, J., Whicker, C. L., & Kim, M. J. (2020). Identifying the Representational Structure of Affect Using fMRI. *Affective Science*, *1*(1), 42–56. https://doi.org/10.1007/s42761-020-00007-9

Matthews, C. M., Thierry, S. M., & Mondloch, C. J. (2020). Recognizing, Discriminating, and

Labeling Emotional Expressions in a Free-Sorting Task: A Developmental Story. *Emotion*. https://doi.org/10.1037/emo0000851

Morningstar, M., Ly, V. Y., Feldman, L., & Dirks, M. A. (2018). Mid-adolescents' and adults' recognition of vocal cues of emotion and social intent: Differences by expression and speaker age. *Journal of Nonverbal Behavior*, *42*(2), 237-251.

Nelson, N. L., & Russell, J. A. (2011). Preschoolers' use of dynamic facial, bodily, and vocal cues to emotion. *Journal of experimental child psychology*, *110*(1), 52-61.

Neta, M., Cantelon, J., Mahoney, C. R., Taylor, H. A., & Davis, F. C. (2017). The impact of uncertain threat on affective bias: Individual differences in response to ambiguity. *Emotion,* 17(8):1137-1143. https://doi.org/10.1037/emo0000349.

Neta, M., Tong, T. T., & Henley, D. J. (2018). It's a matter of time (perspectives): shifting valence responses to emotional ambiguity. *Motivation and Emotion*, *42*(2), 258-266.

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. https://doi.org/10.1038/s41562-017-0238-7

Nook, E. C., & Somerville, L. H. (2019). Emotion concept development from childhood to adulthood. In *Nebraska Symposium on Motivation* (Vol. 66, pp. 11–41). Springer. https://doi.org/10.1007/978-3-030-27473-3_2

Nook, E. C., Stavish, C. M., Sasse, S. F., Lambert, H. K., Mair, P., McLaughlin, K. A., & Somerville, L. H. (2020). Charting the development of emotion comprehension and abstraction from childhood to adulthood using observer-rated and linguistic measures. *Emotion*, *20*(5), 773–792. https://doi.org/10.1037/emo0000609

Norman, G. J., Berntson, G. G., & Cacioppo, J. T. (2014). Emotion, somatovisceral afference, and autonomic regulation. *Emotion Review*, *6*(2), 113-123.

Norris, C. J., Gollan, J., Berntson, G. G., & Cacioppo, J. T. (2010). The current status of research on the structure of evaluative space. *Biological psychology*, *84*(3), 422-436.

Ochsner, K. N., & Phelps, E. (2007). Emerging perspectives on emotion–cognition interactions. *Trends in Cognitive Sciences*, *11*(8), 317-318.

Oksanen, J. (2019). *Vegan: an introduction to ordination*. *R Project*, https://cran.r-project.org/web/packages/vegan/vignettes/intro-vegan.pdf.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*(2), 148-158.

Plate, R. C., Bloomberg, Z., Bolt, D. M., Bechner, A. M., Roeber, B. J., & Pollak, S. D. (2019). Abused children experience high anger exposure. *Frontiers in psychology*, *10*, 440.

Plate, R. C., Wood, A., Woodard, K., & Pollak, S. D. (2019). Probabilistic learning of emotion categories. *Journal of Experimental Psychology: General*, *148*(10), 1814.

Plate, R. C., Woodard, K. & Pollak, S.D. (2021). Statistical learning in an emotional world. Oxford Handbook of Emotion.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665-681.

Plunkett, K. (2011). The role of auditory stimuli in infant categorization. *Infant perception and cognition: Recent advances, emerging theories, and future directions*, 203-221.

Pollak, S. D., Cicchetti, D., Hornung, K., & Reed, A. (2000). Recognizing emotion in faces: developmental effects of child abuse and neglect. *Developmental psychology*, *36*(5), 679.

Pollak, S. D., & Sinha, P. (2002). Effects of early experience on children's recognition of facial

displays of emotion. *Developmental Psychology*, *38*(5), 784.

Pollak, S. D., Camras, L. A., & Cole, P. M. (2019). Progress in understanding the emergence of human emotion. *Developmental Psychology*, *55*(9), 1801.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846-850.

Richie, R., White, B., Bhatia, S., & Hout, M. C. (2020). The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior Research Methods*, *52*(5), 1906-1928.

Ruba, A. L., Johnson, K. M., Harris, L. T., & Wilbourn, M. P. (2017). Developmental changes in infants' categorization of anger and disgust facial expressions. *Developmental Psychology*, *53*(10), 1826-1832.

Ruba, A. L., Meltzoff, A. N., & Repacholi, B. M. (2020). Superordinate categorization of negative facial expressions in infancy: The influence of labels. *Developmental psychology*, *56*(4), 671-685.

Ruba, A. L., & Pollak, S. D. (2020). The development of emotion reasoning in infancy and early childhood. *Annual Review of Developmental Psychology*, *2*, 503-531. https://doi.org/10.1146/annurev-devpsych-060320-102556

Ruba, A. L., & Repacholi, B. M. (2020). Do preverbal infants understand discrete facial expressions of emotion?. *Emotion Review*, *12*(4), 235-250.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social*

psychology, *76*(5), 805.

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

Russell, J. A. (2017). Mixed emotions viewed from the psychological constructionist perspective. *Emotion Review*, *9*(2), 111-117.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, and Psychophysics, 71(6), 1207-1218.* https://doi.org/10.3758/APP.71.6.1207

Sanborn, A. N., Heller, K., Austerweil, J. L., & Chater, N. (2021). REFRESH: A new approach to modeling dimensional biases in perceptual similarity and categorization. *Psychological Review*.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, *51*(4), 1928-1941.

Schwering, S. C., Ghaffari-Nikou, N. M., Zhao, F., Niedenthal, P. M., & MacDonald, M. C. (2021). Exploring the relationship between fiction reading and emotion recognition. *Affective Science*, *2*(2), 178-186.

Shablack, H., & Lindquist, K. A. (2019). The role of language in emotional development. In *Handbook of emotional development* (pp. 451-478). Springer, Cham.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in cognitive sciences*, *7*(6), 246-251.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: a

similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166.

Sloutsky, V. M., & Deng, W. (2019). Categories, concepts, and conceptual development. *Language, cognition and neuroscience*, *34*(10), 1284-1297.

Snyder, H. R., & Munakata, Y. (2010). Becoming self-directed: Abstract representations support endogenous flexibility in children. *Cognition*, *116*(2), 155-167.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33-40.

Srinivasan, R., & Martinez, A. M. (2018). Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2018.2887267

Sutherland, M. R., & Mather, M. (2018). Arousal (but not valence) amplifies the impact of salience. *Cognition and Emotion*, *32*(3), 616-622.

Tillman, K. A., & Barner, D. (2015). Learning the language of time: Children's acquisition of duration words. *Cognitive psychology*, *78*, 57-77.

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. https://doi.org/10.1016/j.jecp.2016.01.005

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. https://doi.org/10.1016/j.jecp.2018.10.007

Vales, C., States, S. L., & Fisher, A. V. (2020). Experience-Driven Semantic Differentiation: Effects of a Naturalistic Experience on Within-and Across-Domain Differentiation in Children. *Child development*, *91*(3), 733-742.

Vales, C., Stevens, P., & Fisher, A. V. (2020). Lumping and splitting: Developmental changes in the structure of children's semantic networks. *Journal of Experimental Child Psychology*, *199*, 104914. https://doi.org/10.1016/j.jecp.2020.104914

Walden, T. A., & Ogan, T. A. (1988). The Development of Social Referencing. *Child Development*, *59*(5), 1230. https://doi.org/10.2307/1130486

Walle, E. A., Reschke, P. J., & Knothe, J. M. (2017). Social referencing: Defining and delineating a basic process of emotion. *Emotion Review*, *9*(3), 245-252.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

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. https://doi.org/10.3758/s13428-012-0314-x

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, *76*, 820-838.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, *29*(3), 257-302.

Waxman, S. R., & Namy, L. L. (1997). Challenging the notion of a thematic preference in young children. *Developmental Psychology*, *33*(3), 555.

Weatherholtz, K., & Jaeger, T. F. (2016). Speech Perception and Generalization Across Talkers and Accents. In *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/ACREFORE/9780199384655.013.95

White, H., Chroust, A., Heck, A., Jubran, R., Galati, A., & Bhatt, R. S. (2019). Categorical perception of facial emotions in infancy. *Infancy*, *24*(2), 139-161.

Wickelmaier, F. (2003). An introduction to MDS. *Sound Quality Research Unit, Aalborg University, Denmark*, *46*(5), 1-26.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Widen, S. C., & Russell, J. A. (2003). A closer look at preschoolers' freely produced labels for facial expressions. *Developmental psychology*, *39*(1), 114.

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, *23*(2), 291–312. https://doi.org/10.1016/j.cogdev.2008.01.002

Widen, S. C., & Russell, J. A. (2010a). The “disgust face” conveys anger to children. *Emotion*, *10*(4), 455.

Widen, S. C., & Russell, J. A. (2010b). Differentiation in Preschooler’s Categories of Emotion. *Emotion*, *10*(5), 651–661. https://doi.org/10.1037/a0019005

Widen, S. C. (2013). Children’s Interpretation of Facial Expressions: The Long Path from Valence-Based to Specific Discrete Categories. *Emotion Review*, *5*(1), 72–77. https://doi.org/10.1177/1754073912451492

Winkler-Rhoades, N., Medin, D., Waxman, S., Woodring, & J., Ross, N. (2010). Naming the animals that come to mind: Effects of culture and experience on category fluency. *Journal of Cognition and Culture*, *10*(1-2), 205-220.

Woodard, K., Plate, R. C., Morningstar, M., Wood, A., & Pollak, S. D. (2021). Categorization of Vocal Emotion Cues Depends on Distributions of Input. *Affective science*, 1-10.

Woodard, K., Plate, R. C., & Pollak, S. D. (2021). Children track probabilistic distributions of

facial cues across individuals. *Journal of Experimental Psychology: General*.

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Current Protocols in Bioinformatics*, *69*(1), https://doi.org/10.1002/cpbi.96

Zemla, J. C., & Austerweil, J. L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational brain & behavior*, *1*(1), 36-58.

Zemla, J. C., Cao, K., Mueller, K. D., & Austerweil, J. L. (2020). SNAFU: The semantic network and fluency utility. *Behavior research methods*, *52*(4), 1681-1699.

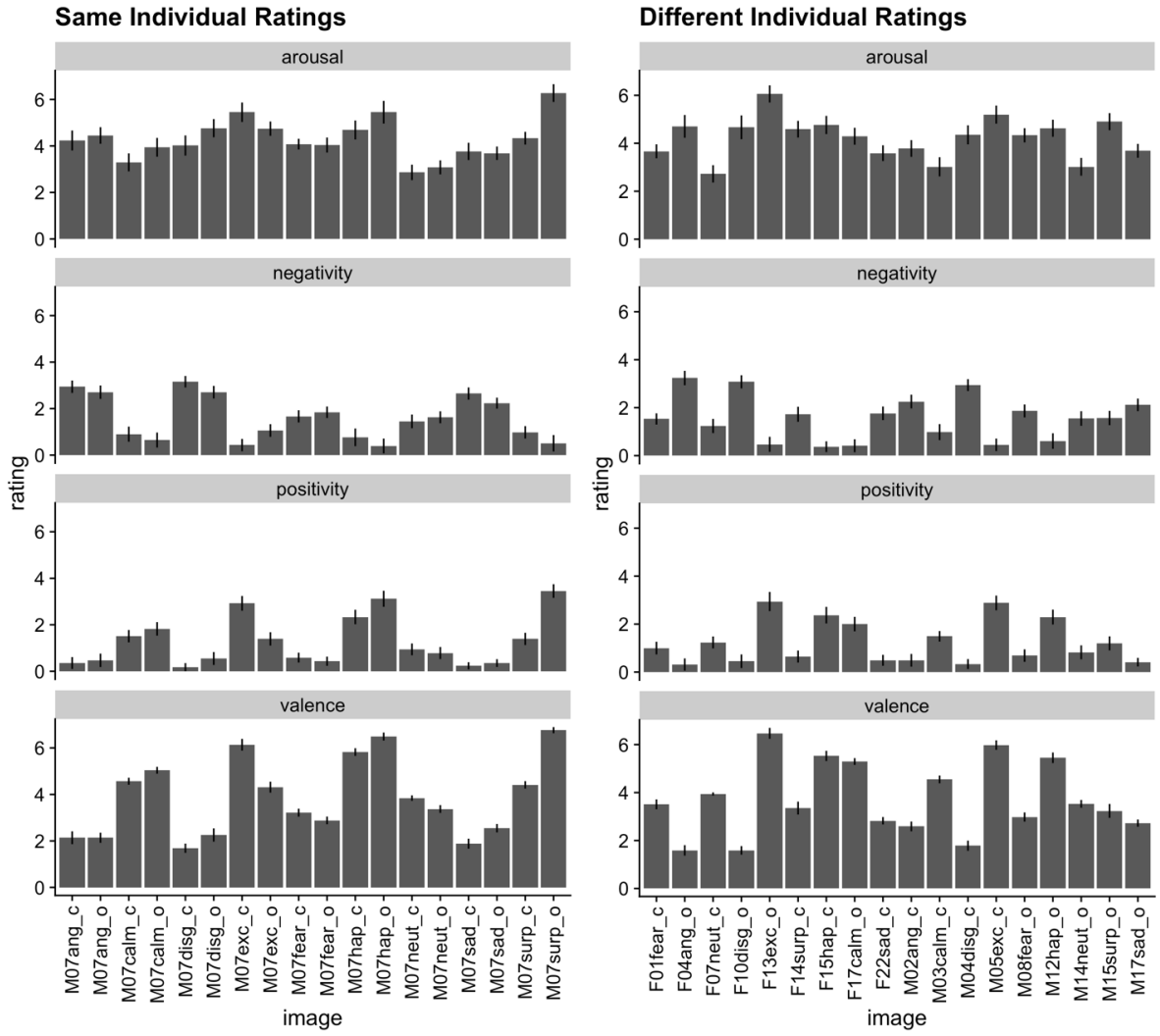
Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K., & Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition & Behavior Reviews*.

Appendix A on Stimulus Ratings

Ratings of bipolar valence, bivariate valence, and arousal were collected for all 36 images from 50 undergraduates who did not complete the sorting tasks. Figure A1 contains detailed information about the average rating for each stimulus. For both sorts, bipolar valence and bivariate valence (positivity and negativity) were highly correlated ($r > 0.85$), while arousal had small to moderate correlations with the other dimensions (see Table A1).

Figure A1

Average Ratings for all 36 Stimuli for the Same and Different Individual Sorts



Note. All stimuli are labeled with the actor's gender (M/F) followed by the actor's number, the emotion displayed, and the mouth position (open = o; closed = c). Error bars represent 95% confidence intervals.

Table A1

Correlations of Stimuli Ratings for the Different Dimensions

	Same Individual				***Different Individuals***			
	Valence	*Arousal*	*Pos.*	*Neg.*	*Valence*	*Arousal*	*Pos.*	*Neg.*
Valence	1.0	0.38	0.93	0.86	1.0	0.17	0.90	0.90
Arousal	0.38	1.0	0.56	0.07	0.17	1.0	0.26	-0.04
Positivity	0.93	0.56	1.0	0.64	0.90	0.26	1.0	0.68
Negativity	0.86	0.07	0.64	1.0	0.90	-0.04	0.68	1.0

Appendix B on Gender

On average, females tend to have better emotion recognition and discrimination than males (e.g., Hall & Gunnery, 2013), a difference already present in early childhood (e.g., Brody, 2000). Many of the differences that emerge in emotion recognition are thought to be due to differences in socialization (e.g., Brody & Hall, 2010). The present study examined whether female and male children (N=107; 48 M, 59 F; mean male age = 4.96, sd_{males} = 1.14; mean female age = 5.05, $sd_{females}$ = 1.07) displayed differences in the use of emotion categories and dimensions to guide their sorting behaviors. To do so, I fit a linear mixed-effects model estimating the average distance between item pairs for children (age in years; centered), the category match for an image pair (same category pair vs. different category pair, centered: same = 0.5, different = -0.5), gender (female = -0.5, male = 0.5) and their interactions with a by-participant random intercept and a by-participant random slope for category match. There were

no differences in males' and females' use of emotion categories to guide sorting behavior (b = 0.004, Wald 95% CI=[-0.02, 0.03], $F(1, 128.22) = 0.08$, $p = .78$).

Next, I looked at children's use of dimensions (valence, arousal) and ran the above model but with the dimension of interest rather than the category match. Again, there were no gender differences in the use of dimensions to guide sorting behavior (bipolar valence: $b = 0.001$, Wald 95% CI =[-.01, 0.01], $F(1, 102.08) = 0.59$, $p = 0.45$; arousal: $b = 0.001$, Wald 95% CI =[-0.01, 0.01], $F(1, 102.00) = .04$, $p = .85$; positivity: $b = -0.002$, Wald 95% CI =[-0.02, 0.01], $F(1, 101.97) = 0.08$, $p = .77$; negativity: $b = 0.003$, Wald 95% CI =[-0.02, 0.02] $F(1, 101.89) = 0.11$, $p = .74$).

Last, I examined whether the verbal fluency task contained any gender differences, as girls sometimes have larger expressive vocabularies. A linear regression model predicting word count using age (centered), gender (F = -0.5, M = .5), and their interaction found that older children produced more animal ($b = 2.26$, $SE = 0.44$, $t = 5.16$, $p < .001$) and emotion words ($b = 1.21$, $SE = 0.19$, $t = 6.28$, $p < .001$), and that boys produced marginally more animal words than girls ($b = 1.83$, $SE = 0.93$, $t = 1.95$, $p = .054$; see Table B1). There were no differences in emotion words produced.

Table B1

Average Word Count by Gender and Condition (N=90)

Gender	**Age**		**Animal Words**		**Emotion Words**	
	Average	*Range*	*Average*	*Range*	*Average*	*Range*
Female	5.08	3.17-6.92	7.73	1-24	3.88	0-7
Male	5.16	3.08-6.92	9.74	2-23	3.87	0-9

Overall, the present study did not find evidence of gender differences in children's similarity judgments of facial expressions or verbal fluency of emotion words. This finding could indicate that the tasks used in the present study are not sensitive measures of gender differences. An alternative (and optimistic!) interpretation could be that the lack of differences is indicative of diminishing gender differences in emotion knowledge due to broader changes in socialization.