Developmental changes in the ability to draw distinctive features of object categories

Anonymous CogSci submission

Abstract

How do children's visual concepts change across childhood, and how might these changes be reflected in their drawings of object categories? Here, we investigate developmental changes in children's ability to emphasize the relevant visual distinctions between object categories in their drawings. We collected a large-scale dataset of children's drawings from a free-standing drawing station in a children's museum, gathering over 13K drawings from children ages 2-10 years. We hypothesized that older children would produce more recognizable drawings, and that this gain in recognizability would not be entirely explained by concurrent visuomotor developments. To evaluate this prediction, we first applied a pretrained deep convolutional neural network model (VGG-19) to extract a high-level feature representation of all drawings. We then assessed the degree to which these drawings were recognizable by training a linear classifier on these features, and we quantified children's concurrent visuomotor development by assessing their ability to accurately trace complex shapes. We found consistent gains in the recognizability of drawings across childhood that were not attributable to children's developing visuomotor abilities. Furthermore, these gains in classification were accompanied by an increase in the distinctiveness between pairs of categories (e.g., dogs vs. rabbits) in this visual feature space. Overall, these results demonstrate that children's drawings include increasingly more distinctive visual features as they grow older.

Keywords: object representations; child development; visual production; deep neural networks

Introduction

Children draw prolifically, providing a rich source of potential insight into their emerging understanding of the world (Kellogg, 1969). Accordingly, drawings have been used to probe developmental change in a wide variety of domains (Fury, Carlson, & Sroufe, 1997; Karmiloff-Smith, 1990; e.g., Piaget, 1929). In particular, drawings have long provided inspiration for scientists investigating how children represent visual concepts (Minsky & Papert, 1972). For example, even when drawing from observation, children tend to include features that are not visible from their vantage point, yet are diagnostic of category membership (e.g., a handle on a mug) (Barrett & Light, 1976; Bremner & Moore, 1984). As children learn the diagnostic properties of objects and how to recognize them, they may express this knowledge in their drawings of these categories. Indeed, children's visual recognition abilities have a protracted developmental trajectory: configural visual processing—the ability to process relationships between object parts (Juttner, Mller, & Rentschler, 2006; Juttner, Wakui, Petters, & Davidoff, 2016)—matures slowly throughout childhood, as does the ability to recognize objects under unusual poses or lighting (Bova et al., 2007).

Inspired by this prior work, our goal is to understand how developmental changes in how children draw visual concepts relate to their representations of these visual concepts. In particular, we hypothesize that children's drawings may become more recognizable as children learn the distinctive features of particular categories that set them apart from other similar objects (see Figure 1). However, this goal raises several methodological challenges to overcome.

First, it requires a principled and generalizable approach to encoding the high-level visual properties of drawings that expose the extent to which they contain category-diagnostic information (Fan, Yamins, & Turk-Browne, 2018). This is in contrast to previous approaches, which have relied upon provisional criteria specific to each study (e.g., handles for mugs) (e.g., Barrett & Light, 1976; Goodenough, 1963), which limited their ability to make detailed predictions on new tasks or datasets. We meet this challenge by capitalizing on recent work validating the use of deep convolutional neural network (DCNN) models as a general basis for measuring the high-level visual information that drives recognition in images, including sparse drawings of objects (Fan et al., 2018; Long, Fan, & Frank, 2018; Yamins et al., 2014). Here, we evaluate whether children include distinctive features in their drawings by asessing whether well these visual features can be used to identify the category (e.g., dog, bird) children were intending to draw.

Second, it requires a large sample of drawings collected under consistent conditions from a wide range of participants to identify robust developmental patterns (e.g., Frank et al., 2017). This is in contrast to the relatively small samples that have characterized classic studies in this domain (Bremner & Moore, 1984; Karmiloff-Smith, 1990). To meet this challenge, we installed a free-standing drawing station in a local science museum, allowing us to collect a large sample of drawings ($N \geq 13205$ drawings) over a large developmental age range (2-10 years) of a variety of object categories (e.g., cup, cat, couch, sheep) under consistent task conditions.

Third, it requires simultaneous and detailed measurement of developmental changes in other cognitive and motor abilities that may influence children's ability to include relevant information in their drawing (Freeman, 1987; Rehrig & Stromswold, 2018). For example, children's developing

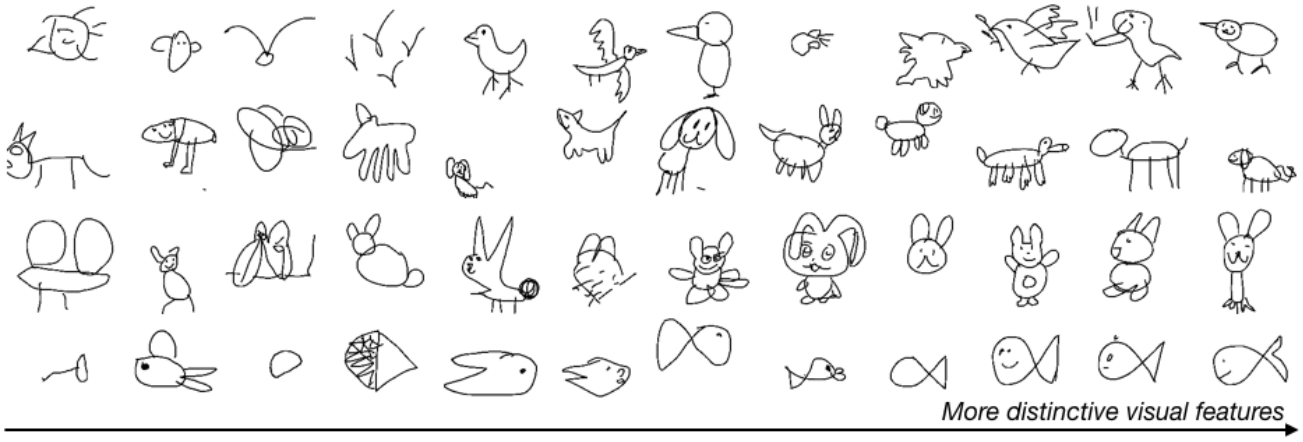

Figure 1: Examples of drawings that have increasingly more distinctive visual features of their categories, making them more easily recognizable. These examples are generated from the results of the classification process outlined below.

visuomotor abilities may limit their ability to include the diagnostic visual features in their drawings. In this paper, we focus on visuomotor control, operationalized as performance on shape tracing and copying tasks, because they share many of the same demands on controlled, visually-guided movement with our main object drawing task. Critically, because we collected both tracings and drawings from every participant in our dataset, we are able to model the contribution of both individual and age-related variation in tracing task performance for explaining how well children produce recognizable drawings.

Methods

Dataset

Drawing Station We installed a drawing station in a local science museum that featured a tablet-based drawing game. Each participant sat in front of a table-mounted touchscreen tablet and drew by moving the tip of their finger across the display. Participants gave consent and indicated their age via checkbox, and no other identifying information was collected; our assumption was that parents would navigate this initial screen for children. To measure fine visuomotor control, each session began with two tracing and one copying trial. On each tracing trial, participants were presented with a shape in the center of the display. The first shape was a simple square, and the second was a more complex star-like shape (see Figure 2). On the subsequent copying trial, participants were presented with a simple shape (square or circle) in the center of the display for 2s, then aimed to copy the shape in the same location it had initially appeared. Next, participants completed up to eight object drawing trials. On each of these trials, participants were verbally cued to draw a particular object category by a video recording of an experimenter (e.g., "What about a dog? Can you draw a dog?"). On all trials, participants had up to 30 seconds to complete their tracing, copy, or drawing. There are 23 common object categories represented in our dataset, which were collected across three bouts of data collection focused on 8 of these objects at a time. These categories were chosen to be familiar to children, to cover a wide range of superordinate categories (e.g., animals, vehicles, manipulable objects) and to vary in the degree to which they are commonly drawn by young children (e.g., trees vs. keys).

Dataset Filtering & Descriptives Given that we could not easily monitor all environmental variables at the drawing station that could impact task engagement (e.g., ambient noise, distraction from other museum visitors), we anticipated the need to develop robust and consistent procedures for data quality assurance. We thus adopted strict screening procedures to ensure that any age-related trends we observed were not due to differences in task compliance across age. Early on, we noticed an unusual degree of sophistication in 2-year-old participants' drawings and suspected that adult caregivers accompanying these children may not have complied with task instructions to let children draw on their own. Thus, in later versions of the drawing game, we surveyed participants to find out whether another child or an adult had also drawn during the session; all drawings where interference was reported were excluded from analyses. Out of these 2685 participants, 700 filled out the survey, and 156 reported interference from another child or adult (5.81%). Raw drawing data (N = 15594 drawings) were then screened for task compliance using a combination of manual and automated procedures (i.e., excluding blank drawings, pure scribbles, and drawings containing words), resulting in the exclusion of 15.3% of all drawings (N=13205 drawings after exclusions). After filtering, we analyzed data from 2443 children who were on average 5.28 years of age (range 2-10 years).

Measuring Tracing Accuracy

We developed an automated procedure for evaluating how accurately participants performed the tracing task, validated against empirical judgments of tracing quality. In subsequent work, we will develop an analogous procedure for evaluating copying task performance. We decompose tracing accuracy into two terms: a shape error term and a spatial error term. Shape error reflects how closely the participant's tracing matched the contours of the target shape; the spatial error reflects how closely the location, size, and orientation of the participant's tracing matched the target shape (see Figure 2).

To compute these error terms, we applied an image registration algorithm, Airlab (Sandkhler, Jud, Andermatt, & Cattin, 2018), to align each tracing to the target shape, yielding an affine transformation matrix minimizing the pixel-wise normalized correlation loss $Loss_{NCC} = -\frac{\sum S \cdot T - \sum E(S)E(T)}{N\sum Var(S)Var(T)}$ between the transformed tracing and the target shape, where N is the number of pixels in both images.

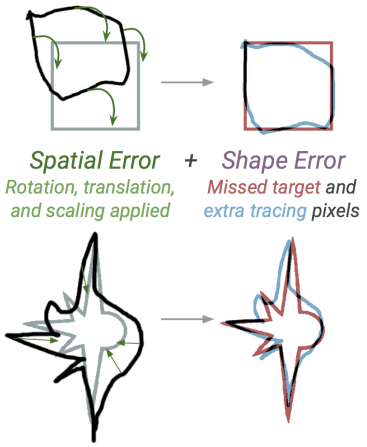

Figure 2: Measurement of tracing task performance reflects both spatial and shape error components. Left: The grey shape is the target; the black shape is the raw tracing. After applying affine image registration, the spatial error reflects the extent of translation, rotation and scaling transformation required to minimize shape error. Right: Shape error reflects how closely the contour of the transformed tracing aligns with the target.

The shape error was defined to be the z-scored cross-correlation loss between the transformed tracing and the target shape. The spatial error was defined to be a combination of three sources of error: location, orientation, and size error, derived by decomposing the affine transformation into translation, rotation, and scaling components. The resulting raw translation, rotation, and scaling errors were then z-scored independently within each spatial error dimension, then summed. This sum was z-scored again to yield the combined spatial error.

Although we assumed that both shape and spatial error should contribute to our measure of tracing task performance, we did not know how much each component contributes to empirical judgments of tracing quality. In order to estimate their relative weights, we collected quality ratings for 1222 tracings (50-80 tracings x 2 shapes x 9 age categories) from adult observers (N=25); 742 tracings were taken from the current dataset. Raters were instructed to evaluate "how well the tracing matches the target shape and is aligned to the position of the target shape" on a 5-point scale. To control for individual variation in the extent to which they used the full range of possible ratings, we z-scored ratings from the same session to map them to the same scale. We then fit a linear mixed-effects model containing shape error, spatial error, their interaction, and shape identity (square vs. star) as predictors of the z-scored empirical ratings. This yielded parameter estimates that could then be used to score each tracing in the remainder of the dataset (N=3242 tracings from 1886 children), and then averaged within session to yield a tracing score for each participant (2245 children completed at least one tracing trial).

Measuring Object Drawing Recognizability

We also developed an automated procedure for evaluating how well participants included category-diagnostic information in their drawings, by examining classification performance on the features extracted by a deep convolutional neural network model.

Visual Encoder To encode the high-level visual features of each sketch, we used the VGG-19 architecture (Simonyan & Zisserman, 2014), a deep convolutional neural network pretrained on Imagenet classification. We used model activations in the second-to-last layer of this network, which contain more explicit representations of object identity than earlier layers (Fan et al., 2018; Long et al., 2018; Yamins et al., 2014). Raw feature representations in this layer consist of flat 4096-dimensional vectors, to which we applied channel-wise normalization.

Logistic Regression Classifier Next, we used these features to train an object category decoder. To avoid any bias due to imbalance in the distribution of drawings over categories (since groups of categories ran at the station for different times), we sampled such that there were an equal number of drawings of each of the 23 categories (N=8694 drawings total). We then trained a 23-way logistic classifier with L2 regularization under leave-one-out cross-validation to estimate the recognizability of every drawing in our dataset.

Predicting Object Drawing Recognizability If children's drawings contain more distinctive features of the drawn categories, then these visual features (estimated via VGG-19) should lead to greater classification accuracy. However, we anticipated that classification accuracy would also vary with children's tracing abilities as well how much time and effort children invested in their drawings; we thus recorded how much time was taken to produce each drawing, how many strokes were drawn, and the proportion of the drawing canvas that was filled. Our main statistical model was then a gener-

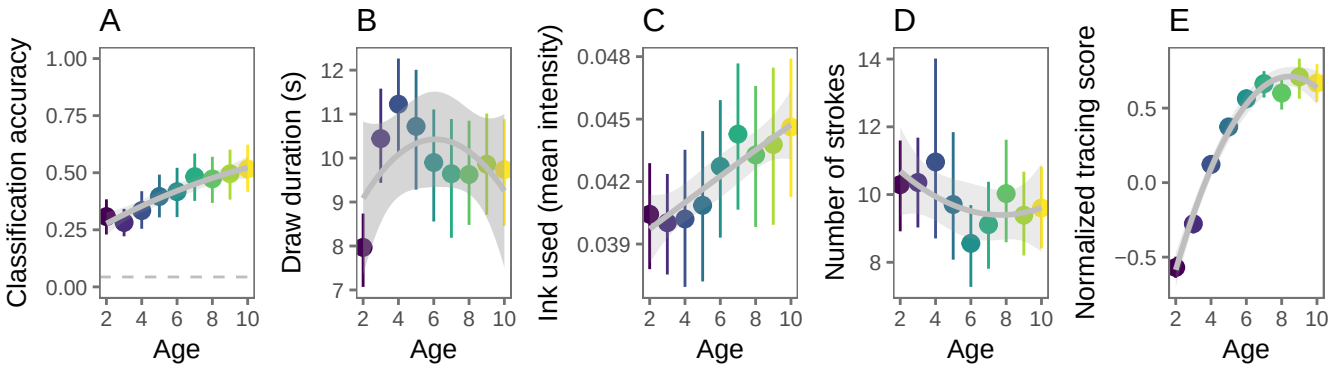

Figure 3: (A) Leave-one-out classification accuracy (grey dotted line indicates chance) (B) the amount of time spent drawing in seconds, (C) the amount of ink used (i.e., mean intensity of the drawings),(D) the number of strokes used, and (E) the average normalized tracing scores are plotted as a function of childrens age.

alized linear mixed-effects model predicting classification accuracy from the category decoder, with scaled age (in years), tracing score (averaged over both trials), and effort cost variables (i.e., time, strokes, ink) modeled as fixed effects, and with random intercepts for each child and object category.

Measuring Category-Distinctiveness We also evaluated how distinct category clusters were at each age by calculating changes in a high-dimensional analogue of d-prime. This distinctiveness metric computes how discriminable two category representations (e.g., bird, rabbit) are by accounting for both the distance between two categories as well as the dispersion within each category. For each pair of categories in each age group, we first computed the Euclidean distance between their respective category centers, defined as the mean feature vector for each category. We then computed the dispersion of each category as the root-mean-squared Euclidean distance of each drawing vector from this category center. By direct analogy with d-prime, we then divide the Euclidean distance between category centers by the quadratic mean of the two category dispersions. Formally, this is specified as: $D_{ij} = \frac{\|\vec{r}_i - \vec{r}_j\|_2}{\sqrt{\frac{1}{2}(s_i^2 + s_j^2)}}$, where D represents the distinctiveness between the ith and jth object categories, $\vec{r}_i$ and $\vec{r}_j$ are their mean feature vectors, $\|\vec{r}_i - \vec{r}_j\|_2$ is the Euclidean distance between them, and where s represents their category dispersions.

Results

Overall, drawing classification accuracy increased with age (see Figure 3A). Our mixed-effects model on drawing classification revealed that this age-related gain held when accounting for task covariates—-the amount of time spent drawing, the number of strokes, and total ink used (see Figure 3B,C,D)—-and for variation across object categories and individual children. All model coefficients can be found in Table 1.

We next examined the relationship between children's ability to trace complex shapes and the subsequent recognizability of their drawings. Tracing abilities increased with age (see Figure 3E) and individual's tracing abilities were good predictors of the recognizability of the drawings they produced. This main effect of tracing ability also held when accounting for effort covariates (number of strokes, time spent drawing, ink used). However, children's tracing abilities did not interact with the age-related gains in classification we observed (see Figure 4): there was no interaction between age and tracing ability and we observed age-related classification gains at each level of tracing ability.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.700	0.274	-2.555	0.011
Tracing	0.312	0.034	9.143	0.000
Age	0.272	0.034	8.075	0.000
Draw Duration	0.139	0.034	4.055	0.000
Avg Intensity	-0.064	0.033	-1.936	0.053
Num. Strokes	-0.035	0.034	-1.019	0.308
Tracing*Age	-0.017	0.028	-0.613	0.540

Table 1: Model coefficients of a GLMM predicting the recognziability of each drawing.

To examine the contributions of these two factors to recognizability, we also fit reduced versions of the full model and examined the marginal R^2 (Nakagawa & Schielzeth, 2013). The fixed effects in a null model without tracing or age (which mainly captures drawing effort) accounted for very little variance (marginal $R^2 = 0.004$). Adding only children's age to the model increased R^2 (marginal $R^2 = 0.037$) as did only adding tracing (marginal $R^2 = 0.04$). Adding both factors without their interaction (marginal $R^2 = 0.05$) had a similar effect to adding both factors and their interaction (marginal $R^2 = 0.05$). However, the random effects of individual items and participants explained a much larger amount of variance than these fixed effects (conditional R^2 for full model = 0.402)

What changes in the feature space might support these increases in classification accuracy across childhood? We expected that increases in classification would be paralleled by

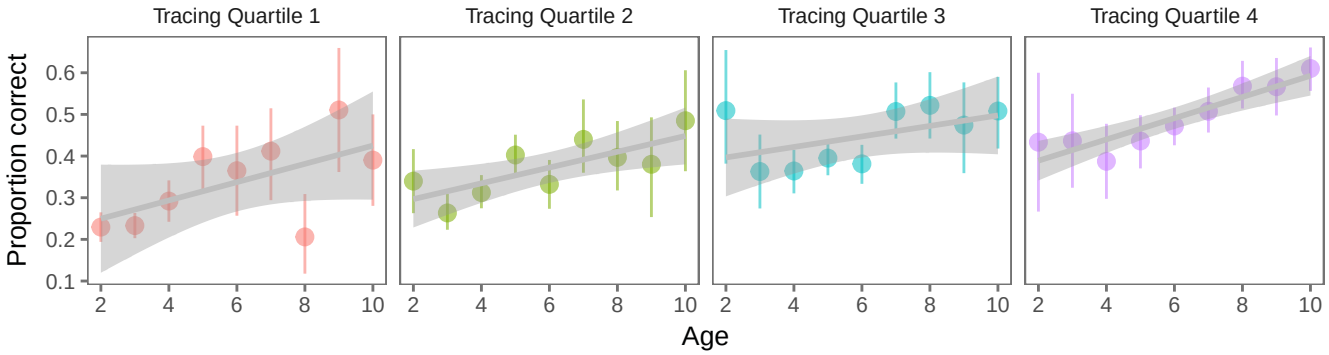

Figure 4: Data are divided into four quantiles based on the distribution of tracing scores in the entire dataset; these divisions represent the data in each panel. In each panel, the average probability assigned to the target class is plotted as a function of childs age. Error bars represent 95% CIs bootstrapped within each age group and subset of tracing scores.

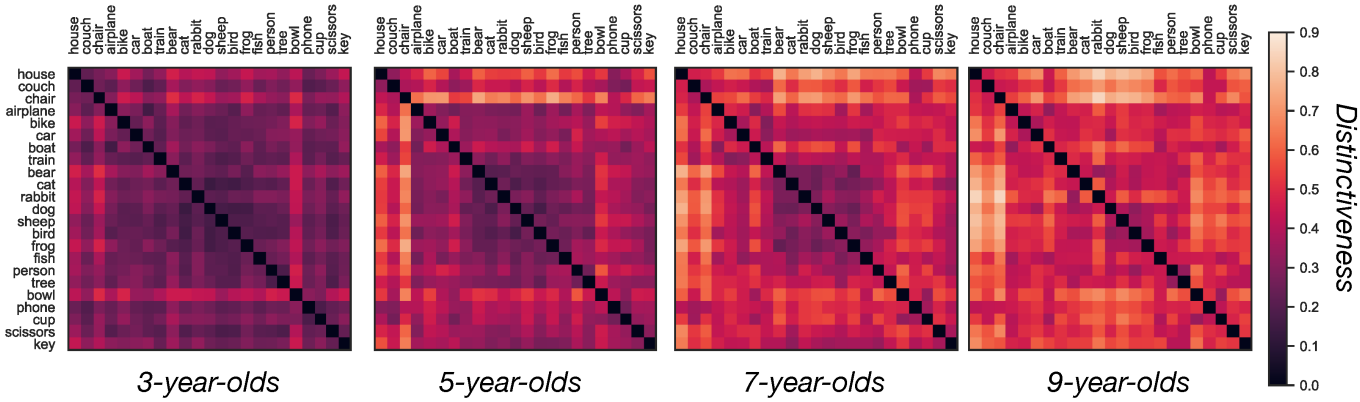

Figure 5: Pairwise category distinctiveness for drawings made by 3-, 5-, 7-,and 9-year-olds; darker (vs. lighter) values represent pairs of categories that have more overlapping (vs. distinctive) representations.

an increase in the distinctiveness of the depicted categories in their high-level visual features. If-in this visual feature space–drawings from the same category became closer together and drawings from different cateogries became farther apart, this would aid classification performance. We thus computed pairwise category "distinctiveness" by evaluating a higher-dimensional analog of d-prime that accounts for both changes in the relative distances between category centers as well as their relative dispersions. Overall, we found an overall increase in the distinctiveness between object categories with age, as well as relative consistency at each age in the pairs of categories that were more or less distinctive (see Figure 5). Taken together, these results suggest that developmental changes in these high-level visual features of children's drawings directly lead to gains in classification accuracy, and that these age-related gains in classification are not entirely explainable by visuomotor developments.

General Discussion

How do children represent different object categories throughout childhood? Drawings are a rich potential source of information about how visual representations change over development. One possibility is that older children's drawings are more recognizable because children are better able to include the distinctive features of particular categories that set them apart from other similar objects. Supporting this hypothesis, the high-level visual features present in children's drawings could be used to estimate the category children were intending to draw, and these classifications became more accurate as children became older. These age-related gains in classification were not entirely explainable by either low-level task covariates (e.g., amount of time spent drawing, average intensity, or number of strokes) or children's tracing abilities. These gains in classification were paralleled by an increase in the distinctiveness between the categories that children drew (see Figure 5).

Taken together, these results suggest that children's drawings contain more distinctive features as they grow older, per-

haps reflecting a change in their internal representations of these categories. While children could simply be learning routines to draw certain categories—perhaps from direct instruction or observation, our results held even when restricted to a subset of very rarely drawn categories (e.g., "couch", "scissors", "key",) providing evidence against a simple version of this idea.

Thus, these descriptive results open the door for future experimental work to link children's drawings to their changing visual concepts. One possibility is that children's drawing of object categories are intimately linked to their visual recognition behaviors: children who produce these more distinctive features in their drawings have finer-grained perceptual representations of these categories that become more specific with age (Pascalis, Haan, & Nelson, 2002). On this account, younger children who tend to not draw these features have more general, lossier visual representations of these categories and show poorer recognition behavior. A second possibility is that when children are asked to draw "a rabbit" they retrieve a list of diagnostic attributes that rabbits possess (e.g., long ears, whiskers). If so, we might observe a relationship between the attributes that children generate when asked to verbally describe "a rabbit" and the features that children include in their drawings.

Overall, we suggest that children's drawings change systematically across development, and that they contain rich information about children's underlying representations of the categories in the world around them. A full understanding of how children's drawings reflect their emerging perceptual and conceptual knowledge will allow a unique and novel perspective on the both the development and the nature of visual concepts—the representations that allow us to easily derive meaning from what we see.

Acknowledgements

We thank the (blinded name of museum) for their collaboration in setting up the drawing station, and we thank members of (blinded name of lab) for valuable feedback. This work was funded by an NSF SPRF-FR Grant #XXXXXX to XXX and a Jacobs Foundation Fellowship to XXX.

References

Barrett, M., & Light, P. (1976). Symbolism and intellectual realism in children's drawings. *British Journal of Educational Psychology*, *46*(2), 198–202.

Bova, S. M., Fazzi, E., Giovenzana, A., Montomoli, C., Signorini, S. G., Zoppello, M., & Lanzi, G. (2007). The development of visual object recognition in school-age children. *Developmental Neuropsychology*, *31*(1), 79–102.

Bremner, J. G., & Moore, S. (1984). Prior visual inspection and object naming: Two factors that enhance hidden feature inclusion in young children's drawings. *British Journal of Developmental Psychology*, *2*(4), 371–376.

Fan, J., Yamins, D., & Turk-Browne, N. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *0*(0). http://doi.org/10.1111/cogs.12676

Frank, M., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... others. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

Freeman, N. H. (1987). Current problems in the development of representational picture-production. *Archives de Psychologie*.

Fury, G., Carlson, E. A., & Sroufe, A. (1997). Children's representations of attachment relationships in family drawings. *Child Development*, *68*(6), 1154–1164.

Goodenough, F. L. (1963). *Goodenough-harris drawing test*. Harcourt Brace Jovanovich New York.

Juttner, M., Mller, A., & Rentschler, I. (2006). A developmental dissociation of view-dependent and view-invariant object recognition in adolescence. *Behavioural Brain Research*, *175*(2), 420–424.

Juttner, M., Wakui, E., Petters, D., & Davidoff, J. (2016). Developmental commonalities between object and face recognition in adolescence. *Frontiers in Psychology*, *7*.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*(1), 57–83.

Kellogg, R. (1969). *Analyzing children's art*. National Press Books Palo Alto, CA.

Long, B., Fan, J., & Frank, M. (2018). Drawings as a window into the development of object category representations. *Journal of Vision*, *18*(10), 398–398.

Minsky, M., & Papert, S. A. (1972). Artificial intelligence progress report.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Pascalis, O., Haan, M. de, & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? *Science*, *296*(5571), 1321–1323.

Piaget, J. (1929). The child's concept of the world. *Londres, Routldge & Kegan Paul*.

Rehrig, G., & Stromswold, K. (2018). What does the dap: IQ measure?: Drawing comparisons between drawing performance and developmental assessments. *The Journal of Genetic Psychology*, *179*(1), 9–18.

Sandkhler, R., Jud, C., Andermatt, S., & Cattin, P. C. (2018). AirLab: Autograd image registration laboratory. *ArXiv Preprint ArXiv:1806.09907*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.