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

Bria Long
Department of Psychology
Stanford University
450 Serra Mall
Stanford, CA 94305
bria@stanford.edu

Judith E. Fan
Department of Psychology
UC San Diego
9500 Gilman Drive
La Jolla, CA 92093
jefan@ucsd.edu

Zixian Chai
Department of Psychology
Stanford University
450 Serra Mall
Stanford, CA 94305
zchai14@stanford.edu

Michael C. Frank
Department of Psychology
Stanford University
450 Serra Mall
Stanford, CA 94305
mcfrank@stanford.edu

Abstract

How do children's visual concepts change across childhood, and how might these changes be reflected in their drawings? Here we investigate developmental changes in children's ability to emphasize the relevant visual distinctions between object categories in their drawings. We collected over 13K drawings from children aged 2-10 years via a free-standing drawing station in a children's museum. We hypothesized that older children would produce more recognizable drawings, and that this gain in recognizability would not be entirely explained by concurrent development in visuomotor control. To measure recognizability, we applied a pretrained deep convolutional neural network model to extract a high-level feature representation of all drawings, and then trained a multi-way linear classifier on these features. To measure visuomotor control, we developed an automated procedure to measure their ability to accurately trace complex shapes. We found consistent gains in the recognizability of drawings across ages that were not fully explained by children's ability to accurately trace complex shapes. Furthermore, these gains were accompanied by an increase in how distinct different object categories were in feature space. Overall, these results demonstrate that children's drawings include 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, Muller, & Rentschler, 2006; Juttner, Wakui, Petters, & Davidoff, 2016)—may mature 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 the relationship between developmental changes in how children draw visual concepts and their representations of these visual concepts. In particular, we hypothesize that children's drawings become more recognizable in part because children learn the distinctive features of categories that set them apart from other similar categories (Figure 1). If so, we would expect an increase in the distinctiveness of children's drawings across childhood that is not explained by improvements in children's visuomotor ability. However, this goal poses 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 approach stands 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. Recently, deep convolutional neural network (DCNN) models that have been trained on challenging object recognition tasks have been shown to extract high-level visual information from images (Yamins et al., 2014). As these models have been directly optimized to recognize objects in photographs, features in higher layers of these networks represent high-level visual information that is important for distinguishing between object categories. We thus meet this challenge by capitalizing on prior work validating the use of these higher-layer features to analyze the high-level visual information in drawings (Fan et al., 2018; Long, Fan, & Frank, 2018). In particular, we investigate the extent to which children include distinctive features in their drawings by assessing how well these visual features can be used to identify the category (e.g., dog, bird) that 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., M. 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 = 13205 drawings) of 23 object categories

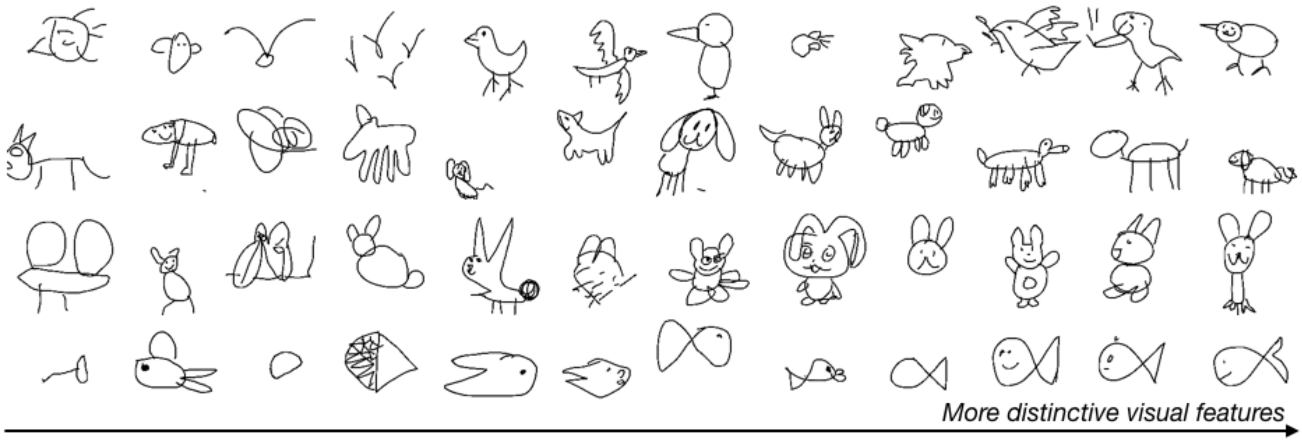

Figure 1: Examples of drawings that have increasingly more distinctive visual features of their categories, making them more easily recognizable. These examples are generated from the results of the classification process outlined below.

over a wide developmental age range (i.e., 2-10 years) 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 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 tasks, because they share many of the same demands on controlled, visually-guided movement with our primary 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.

In sum, our paper provides an advance over our prior work investigating developmental change in drawing behavior (Long et al., 2018) in three ways: first, we build a free-standing drawing station to continually crowdsource children's drawings under consistent conditions, enabling the collection of a substantially larger dataset; second, we exploit this larger dataset to characterize the category-level distinctiveness inherent to children's drawings across a wide range of ages; and third, we develop an automated procedure for analyzing concurrent changes in visuomotor control using a tracing task.

Methods

Dataset

Drawing Station We installed a drawing station that featured a tablet-based drawing game in a local science museum. 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 (in years 2-10 or adult) via checkboxes 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 trials, followed by a 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 (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, which then disappeared. They then were asked 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-yearold 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

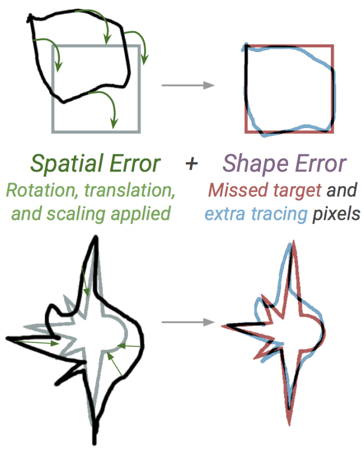

Figure 2: Measurement of tracing task performance reflects both spatial and shape error components. Left: The grey shape is the target; the black shape is the raw tracing. After applying affine image registration, the spatial error reflects the extent of translation, rotation, and scaling transformation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

We developed an automated procedure for evaluating how accurately participants performed the tracing task, validated against empirical judgments of tracing quality. We decompose tracing accuracy into two components: a shape error component and a spatial error component. 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 (Figure 2).

To compute these error components, 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 that minimized the pixel-wise correlation distance between the aligned tracing, T, and the target shape, S: $Loss_{NCC} = -\frac{\sum S \cdot T - \sum E(S)E(T)}{N \sum Var(S)Var(T)}$, where N is the number of pixels in both images.

The shape error was defined by the final correlation distance between the aligned tracing and the target shape. The spatial error was defined by the magnitude of three distinct error terms: location, orientation, and size error, derived by decomposing the affine transformation matrix above into translation, rotation, and scaling components, respectively. In sum, this procedure yielded four error values for each tracing: one value representing the shape error (i.e., the pixel-wise correlation distance) and three values representing the spatial error (i.e., magnitude of translation, rotation, scaling components).

Although we assumed that both shape and spatial error terms should contribute to our measure of tracing task performance, we did not know how much weight to assign to each component to best predict empirical judgments of tracing quality. In order to estimate these weights, we collected quality ratings from adult observers (N=70) for 1325 tracings (i.e., 50-80 tracings per shape per age), each of which was rated 1-5 times. 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.

We fit an ordinal regression mixed-effects model to predict these 5-point ratings, which contained correlation distance, translation, rotation, scaling, and shape identity (square vs. star) as predictors, with random intercepts for rater. This model yielded parameter estimates that could then be used to score each tracing in the remainder of the dataset (N=3242 tracings from 1886 children). We averaged scores within session to yield a single 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 pre-trained 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 cate-

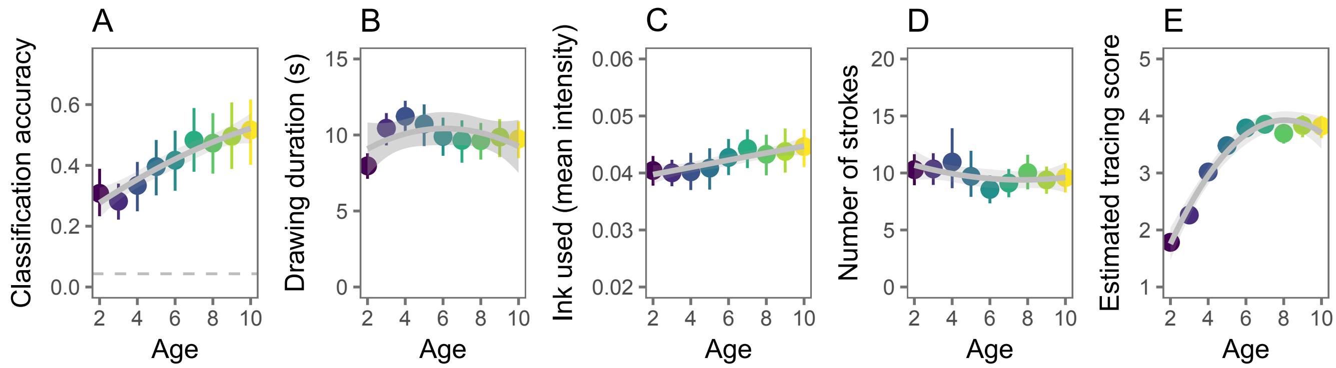

Figure 3: (A) Leave-one-out classification accuracy (grey dotted line indicates chance) (B) the amount of time spent drawing in seconds, (C) the amount of ink used (i.e., mean intensity of the drawings), (D) the number of strokes used, and (E) the average estimated tracing scores are plotted as a function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

gories (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 features that are diagnostic of the drawn categories, then these visual features (estimated via VGG-19) should lead to greater classification accuracy. However, we anticipated that classification accuracy may 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 generalized 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 To investigate changes in the underlying feature representation of children's drawings that may help explain variation in classification accuracy, we computed a measure of pairwise category distinctiveness D_{ij} for each pair of categories i, j within each age. This metric is a higher-dimensional analog of d-prime that incorporates both the distance between each pair of categories as well as the dispersion within each category. We first computed the category centers as the mean feature vector for each category, $\vec{r}_i$ and $\vec{r}_j$. The distance between each pair of categories i, j was then taken as the Euclidean distance between their category centers, $\|\vec{r}_i - \vec{r}_j\|_2$. The dispersion for each category was computed as the root-mean-squared Euclidean distance of each individual drawing vector from the category center vector $\vec{r}$ and is expressed as s. By direct analogy with d-prime, we compute the distinctiveness D_{ij} of each pair of categories i, j by dividing the Euclidean distance between category centers by the quadratic mean of the two category dispersions, $D_{ij} = \frac{\|\vec{r}_i - \vec{r}_j\|_2}{\sqrt{\frac{1}{2}(s_i^2 + s_j^2)}}$.

Results

Overall, drawing classification accuracy increased with age (Figure 3A), validating our basic expectation that older children's drawings would be more recognizable. 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 (Figure 3B,C,D)—and for variation across object categories and individual children. All model coefficients can be found in Table 1.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.714	0.274	-2.606	0.009
Tracing	0.311	0.034	9.141	0.000
Age	0.282	0.033	8.499	0.000
Draw Duration	0.136	0.034	3.976	0.000
Avg Intensity	-0.064	0.033	-1.910	0.056
Num. Strokes	-0.034	0.034	-1.009	0.313
Tracing*Age	0.011	0.029	0.357	0.721

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

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 (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 (Figure 4) and we observed age-related classification gains at each level of tracing ability.

To examine the contributions of age and tracing ability to

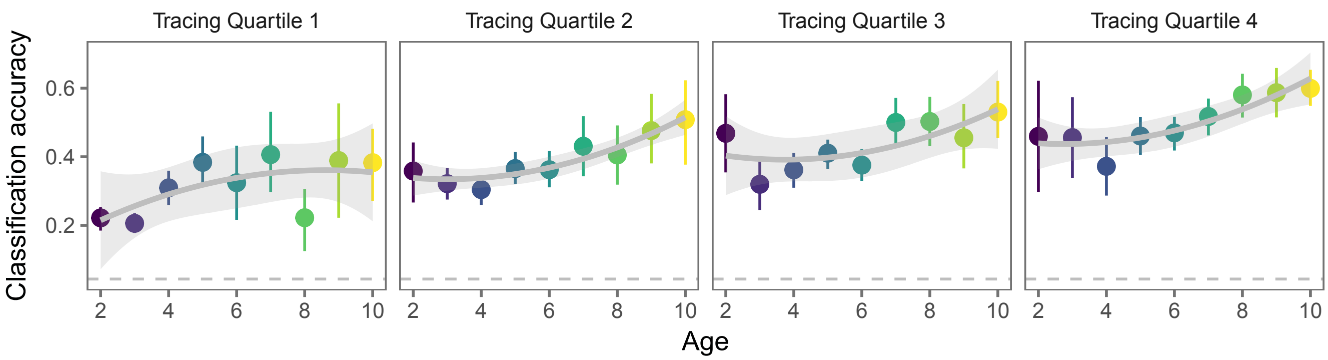

Figure 4: Data are divided into four quantiles based on the distribution of tracing scores in the entire dataset; these divisions represent the data in each panel. In each panel, the average classification accuracy is plotted as a function of childrens age. Error bars represent 95% CIs bootstrapped within each age group and subset of tracing scores; grey dotted lines indicate chance.

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 = 0.039). 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). Attesting to the immense variability between individuals and categories, adding random effects (and many more parameters) accounted for a much larger amount of variance (conditional R^2 for full model = 0.403). Finally, as we had many more younger participants in our dataset, we also repeated these analyses with a subset of the dataset that was balanced across both children's age and category (N=2691 drawings), and found the same pattern of results.

These age-related changes in classification accuracy show that the underlying feature representations of older children's drawings were more linearly discriminable. This finding led us to investigate a potential source of this enhanced discriminability: that drawings from different categories were spread further apart in feature space, while drawings within a category were clustered closer together. To evaluate this possibility, we used a measure of pairwise category distinctiveness D_{ij} that accounts for both the distance between each pair of categories, as well as the dispersion within each category. We found that category distinctiveness increased consistently with age (Figure 5).

Taken together, these results reveal developmental changes in how well children are able to emphasize the relevant distinctions between object categories in their drawings that thereby support recognition. Moreover, they show that these age-related gains in classification are not entirely explained by concurrent development in visuomotor control.

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 diagnostic features of particular categories that distinguish them 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 effort covariates (e.g., amount of time spent drawing, average intensity, or number of strokes) or children's tracing abilities. In addition, these gains in classification were paralleled by an increase in the distinctiveness between the categories that children drew (Figure 5).

Taken together, these results suggest that children's drawings contain more distinctive features as they grow older, perhaps 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) arguing against a simple version of this idea.

Nonetheless, there are limitations on the generalizability of these findings due to the nature of our dataset. First, while this dataset is large and samples a heterogenous population, all drawings were collected at a single geographical location, limiting the generalizability of these results to children from other diverse cultural or socioeconomic backgrounds. Second, while we imposed strong filtering requirements on the dataset, we were not present while the children were drawing and thus cannot be sure that we've eliminated all sources of noise or interference. At the same time, additional interference would only generate extra noise in our data rather than the observed age-related trends. In any case, these correlational results call for validation in more carefully controlled

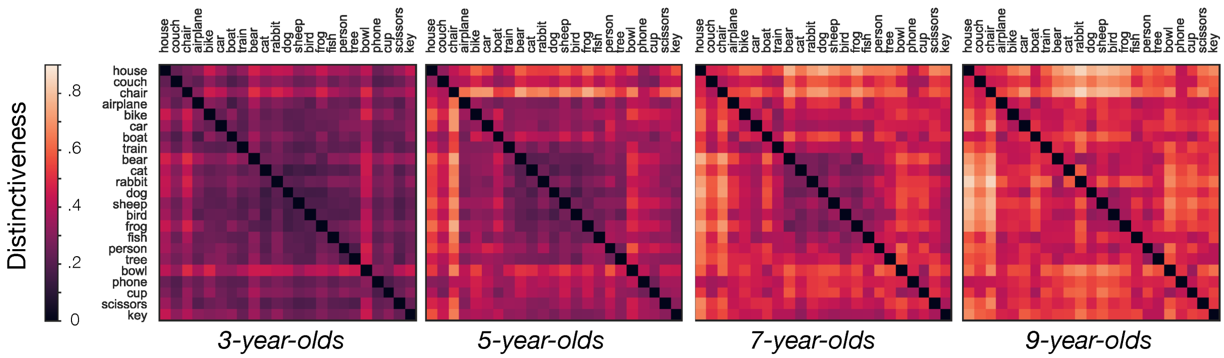

Figure 5: Pairwise category distinctiveness for drawings made by 3-, 5-, 7-,and 9-year-olds; darker (vs. lighter) values represent pairs of categories that have more overlapping (vs. distinctive) representations.

contexts and across more diverse populations.

Furthermore, they open the door for future empirical work to establish causal links between children's drawing behavior and their changing internal representation of visual concepts. For example, it would be valuable to explore the extent to which a child's ability to include the most distinctive visual features in their drawings of object categories predicts their ability to perceptually discriminate those object categories. Another promising direction would be to investigate the relationship between children's general ability to retrieve relevant information from semantic memory (e.g., that a rabbit has long ears and whiskers), and their ability to produce recognizable drawings of those categories. Insofar as such retrieval mechanisms are engaged during drawing production, developmental changes in semantic memory systems may also explain an important portion of the age-related variation in drawing behavior.

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 San Jose Children's Discovery Museum for their collaboration and for hosting the drawing station where these data were collected. We are also grateful to members of the Stanford Language and Cognition lab for their feedback. This work was funded by an NSF SPRF-FR Grant #1714726 to BLL and a Jacobs Foundation Fellowship to MCF.

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).

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., Muller, 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. C. (2018). Drawings as a window into developmental changes in object representations. In *Proceedings of the 40th annual meeting of the cognitive* *science society*.

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.

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.