Parallel developmental changes in children's drawing and recognition of visual concepts

Bria Long[1], Judith E. Fan[2], Zixian Chai[1], Michael C. Frank[1]

[1]Department of Psychology, Stanford University

[2]Department of Psychology, University of California San Diego

Abstract

To what extent do visual concepts of *dogs, cars,* and *clocks* change across childhood? We hypothesized that as children progressively learn which features best distinguish visual concepts from one another, they also improve their ability to connect this knowledge with external representations. To examine this possibility, we investigated developmental changes in children's ability to produce and recognize drawings of common object categories. First, we recruited children aged 2-10 years to produce drawings of 48 categories via a free-standing kiosk in a children's museum, and we measured how recognizable these >37K drawings were using a deep convolutional neural network model of object recognition. Second, we recruited other children across the same age range to identify the drawn category in a subset of these drawings via "guessing games" at the same kiosk. We found consistent developmental gains in both children's ability to include diagnostic visual features in their drawings and in children's ability to use these features when recognizing other children's drawings. Our results suggest that children's ability to connect internal and external representations of visual concepts improves gradually across childhood and imply that developmental trajectories of visual concept learning may be more protracted than previously thought.

Parallel developmental changes in children's drawing and recognition of visual concepts

Introduction

What makes a rabbit look like a rabbit – or a rabbit look different from a cat? As adults, these visual concepts are seamlessly integrated into our experience of the visual world. Toiling in the background, our visual system connects incoming patterns of light with knowledge about what things look like, rapidly inferring the category membership of the objects we see. This flexible, generative visual knowledge also allows us to conceive of and create innumerable new exemplars of visual concepts, communicating about the contents of our mind's eye (Gregory, 1973).

A common view is that visual concepts mature relatively early in life. Supporting this view, even very young infants can extract perceptual commonalities (Quinn & Eimas, 1996) that can distinguish between basic-level categories (e.g., *dog* vs. *cat*). Infants also show a remarkable ability to generalize between different kinds of visual representations: 5-month-olds recognize the correspondence between line drawings and photographs of faces (DeLoache, Strauss, & Maynard, 1979), and when 18-month-olds are taught the name of a novel object depicted in a line drawing, they extend this label to its real-life counterpart (Allen Preissler & Carey, 2004). As young children learn to pair labels (e.g., "rabbit") with their visual experience of objects belonging to different categories — as stylized depictions, toys, or real-life exemplars — this active process may create new features relevant for distinguishing visual categories from one another (for review, see Schyns, Goldstone, & Thibaut, 1998). And by their second birthday, children extend visual concepts appropriately after experiencing one or few exemplars of a novel category (i.e., "one-shot" learning) (Carey & Bartlett, 1978; Pereira & Smith, 2009; Soja, Carey, & Spelke, 1991) and identify abstract exemplars as belonging to those categories (Pereira & Smith, 2009).

However, mounting evidence suggests that children's visual recognition abilities have a relatively extended developmental trajectory throughout middle childhood (for reviews, see Juttner, Wakui, Petters, & Davidoff, 2016; Nishimura, Scherf, & Behrmann, 2009). For example, children become steadily better at discriminating between

perceptually similar exemplars of scenes, objects, bodies, and faces from 5-10 years of age (Weigelt et al., 2014) and increasingly skilled at recognizing objects presented in unusual poses or 3D rotations, reaching adult-like levels only in adolescence (Bova et al., 2007; Dekker, Mareschal, Sereno, & Johnson, 2011; Nishimura, Scherf, Zachariou, Tarr, & Behrmann, 2015). These improvements may partly be because children increasingly attend to the relationship between object parts and features as they approach adolescence (Juttner, Muller, & Rentschler, 2006; Juttner et al., 2016; Mash, 2006). In turn, improvements in children's recognition abilities are reflected in changes in how visual cortex encodes different objects and scenes (Balas & Saville, 2020; Cohen et al., 2019; Dekker et al., 2011; Gomez, Natu, Jeska, Barnett, & Grill-Spector, 2018; Kersey, Clark, Lussier, Mahon, & Cantlon, 2015; Nishimura et al., 2015). For example, children's ability to discriminate between similar faces is correlated with the sensitivity of corresponding face-selective regions to these particular faces (Natu et al., 2016).

Examination of children's drawings also suggests that visual concepts may change throughout childhood (Fury, Carlson, & Sroufe, 1997; Karmiloff-Smith, 1990; Kellogg, 1969; Piaget, 1929). For example, there appear to be dramatic changes in how children encode diagnostic visual information in their drawings across age; younger children (4-5 years) tend to include fewer cues in their drawings to differentiate between target concepts (e.g., *adult* vs. *child*) than older children, who enrich their drawings with more diagnostic part (Sitton & Light, 1992) and relational (Light & Simmons, 1983) information. Furthermore, prior work suggests that children's drawings reflect what they know about objects – and that new experience with these objects in an experimental session can change what children draw. 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), and either haptic or visual experience with an object tends to exaggerate this effect (Bremner & Moore, 1984).

Here, we test a specific hypothesis about the nature of these changes in visual concepts: the idea that children's visual concepts change as they gradually learn to

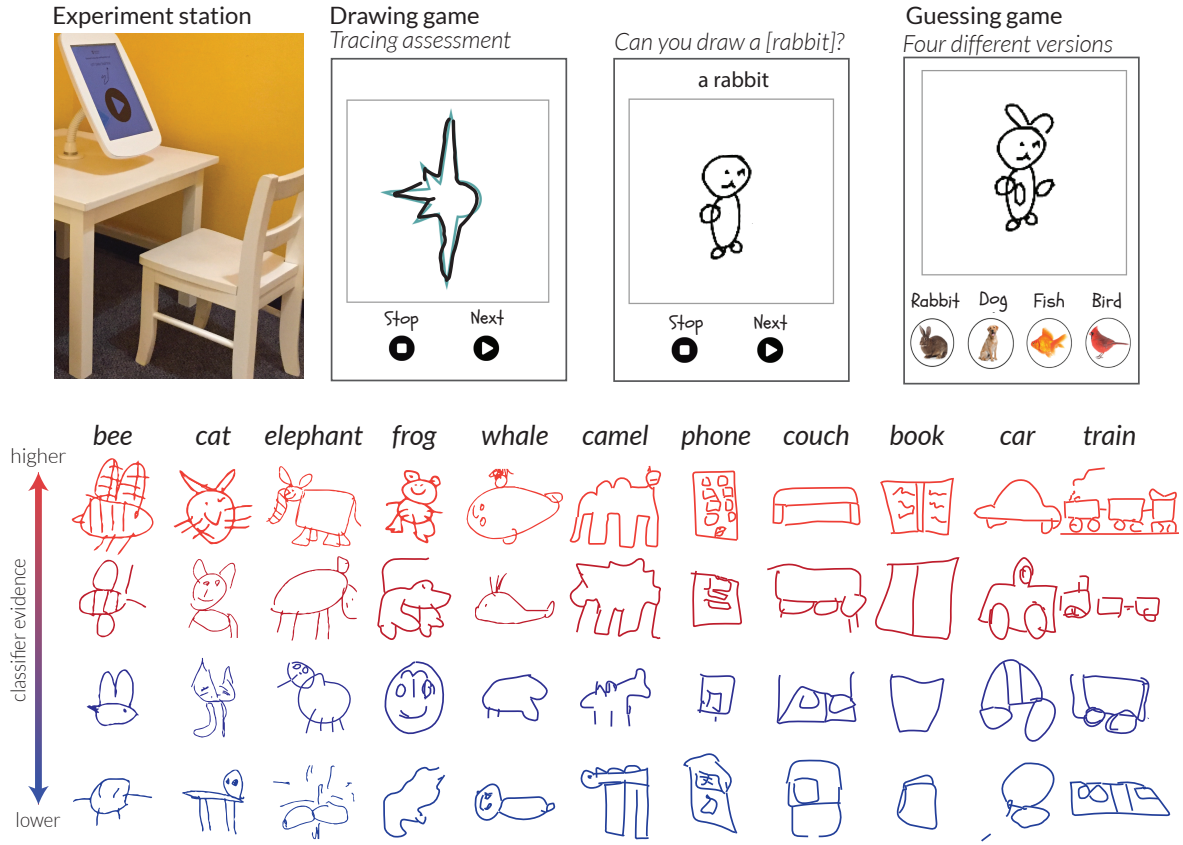

Figure 1. Kiosk setup where children participated and illustration of the tracing trials, drawing station, and guessing game. (Bottom) Example drawings from several categories: redder drawings contain more diagnostic visual features (i.e. classifier evidence, see *Methods*).

identify which features are most diagnostic of each concept. We further hypothesize that changes in how well children can recognize visual concepts are accompanied by parallel changes in how well children can generate recognizable visual representations of them by drawing. To test these predictions, we examine changes in the features that children produce when drawing and in the features that they rely on when recognizing each other's drawings. Our prediction is that children become more sensitive to the distinctive features of visual concepts in both tasks across development.

Alternatively, children's internal visual concepts may not change substantially over childhood. On this account, improvements in children's ability to draw various concepts may solely reflect improvements in their general ability to plan and control

their motor movements (Freeman, 1987; Rehrig & Stromswold, 2018), or their specific experiences drawing particular concepts (Cohn, 2012; Willats, 2006). But on this alternative account, children should not improve in their ability to recognize line drawings of visual concepts, in keeping with accounts of early pictorial competence.

Though the hypothesis that children's visual knowledge has a protracted developmental trajectory has been articulated frequently (e.g. Bova et al., 2007; Sitton & Light, 1992; Sloutsky & Fisher, 2011), testing it has previously been impossible. Until recently there has been no agreed upon or generalizable way to quantify the distinctive, high-level visual features in a given image. In prior work, scientists had to intuit what these distinctive visual features could be (e.g., handles for *mugs*) and then optimize a small set of stimuli and tasks to detect hypothesized changes (e.g. Barrett & Light, 1976; Goodenough, 1963), relying heavily on manual annotations. These methodological obstacles limited the breadth of their conclusions.

To overcome these challenges, here we collected a large, digital sample of children's drawings over childhood for many different categories. We also capitalized on recent improvements in computer vision to analyze this dataset. Recent work has now validated 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, Yamins, & Turk-Browne, 2018; Long, Fan, & Frank, 2018). Activation patterns in higher layers of these models predict adult perceptual judgments of shape similarity (Kubilius, Bracci, & Op de Beeck, 2016), neural population responses to categories throughout object-selective cortex (Yamins et al., 2014), and similarity relationships between sparse drawings and photographs of categories (Fan et al., 2018). Furthermore, unlike human adults, these models have no knowledge of drawing conventions (i.e. how one might typically draw a bird or a fish) and do not incorporate abstract semantic features into their similarity judgements (Hebart, Zheng, Pereira, & Baker, 2020). This new tool makes it possible to assess developmental changes in the visual features of children's drawings. We use these methodological advances to quantify changes in how children produce and recognize

drawings of object categories over development, testing the hypothesis that children's visual concepts change over childhood.

Results

Development of drawing production

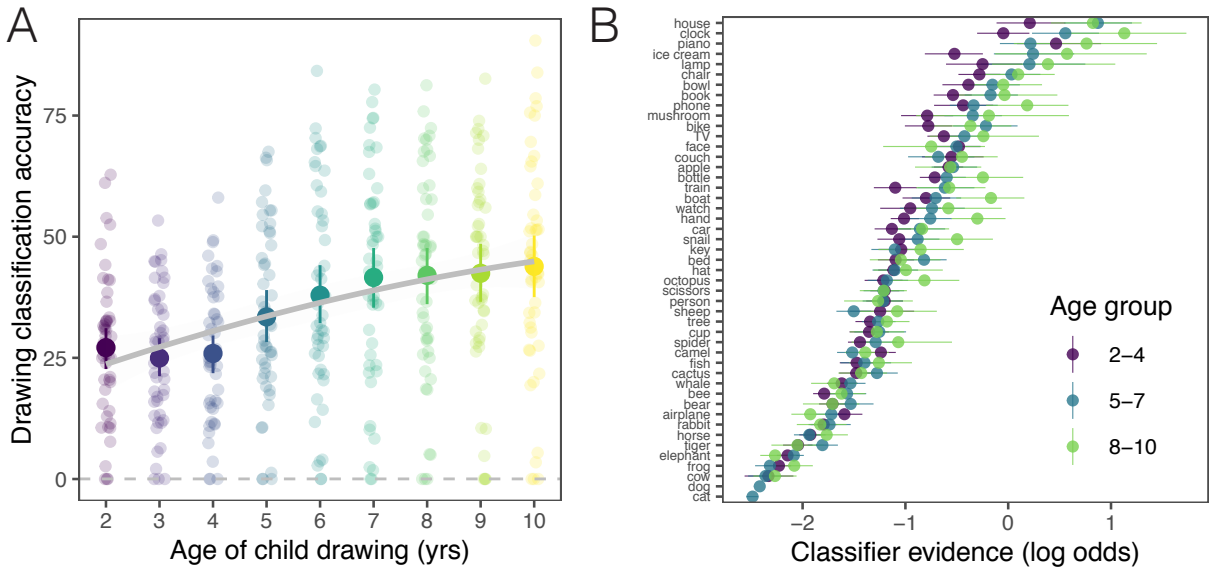

Figure 2. (A). Each dot represents the proportion of drawings that were correctly classified in a given category; chance line represents 1/48 (number of categories in the dataset). (B). Log odds probabilities (i.e. classifier evidence) for a subset of *correctly* recognized drawings, binned by the age group of the child who produced the drawings. Categories are ordered by average log odds probabilities for each category in descending order. Error bars represent bootstrapped 95% confidence intervals in both plots.

A free-standing kiosk designed to be navigable by children was installed at a local science museum (see Figure 1A). Children used a touch-screen tablet to produce their drawings. As children's visuomotor abilities may limit their ability to include the relevant visual features in their drawings (Freeman, 1987; Rehrig & Stromswold, 2018), we also included shape tracing trials to measure children's tracing skills (see Figure 1B). After completing these tracing trails, children were prompted to draw different object categories via verbal prompts. These categories were selected to include both animals and inanimate objects as well as categories that are both commonly (e.g., *cup, face,*

cat) and rarely drawn (e.g., *octopus, piano, camel*) by children. The final, filtered dataset contained 37,770 drawings of 48 categories from N=8084 children who were on average 5.4 years of age (range 2–10 years old; see *Methods* for more details).

To examine changes in the children's drawings across development, we analyzed the degree to which high-level visual features of each drawing could be used to decode the category that children were intending to draw (e.g., *dog*). We obtained these features using a deep convolutional neural network pre-trained on Imagenet classification with a VGG-19 architecture (Simonyan & Zisserman, 2014). Activations for each sketch were taken from the second-to-last layer of this model as prior work has shown that these activations correspond to the visual features that enable basic-level recognition (e.g., *cat* vs. *dog*) in both sketches and photographs (Fan et al., 2018; Yamins et al., 2014). These features were used to train logistic-regression classifiers to predict which of the 48 categories children were asked to draw (e.g., *couch*, *chair*) for sets of held-out drawings (see *Methods*). For every drawing, this procedure thus yielded (1) a binary classification score, indicating whether a given drawing contained the visual features that enabled basic-level recognition, and (2) a probability score for each of the 48 categories, capturing the degree to which a given drawing contained the visual features relevant to that category.

Drawings become more recognizable across childhood. Overall, we found that drawing classification accuracy increased steadily with children's age (see Figure 2A, Table 1), validating the basic expectation that older children's drawings contain visual features that make them more recognizable. One potential explanation for this increase it is driven by specific categories that children tend to draw most frequently (e.g., *trees*, *people*, *dogs*). This account predicts that these frequently drawn categories would show the strongest developmental trends. To test this account, we estimated how frequently children drew each of the categories in the dataset by asking parents to report how often their child tends to draw each category (N=50 parents of children aged 3-10 years, *Methods*). We found little evidence that drawings of more frequently drawn categories were more recognizable or showed stronger developmental trends;

there was neither a main effect of drawing frequency on classification accuracy nor an interaction with age in a generalized linear mixed effect model (see Table 1). In fact, many infrequently drawn categories (e.g. *mushroom*) had relatively high classification accuracy while some frequently drawn categories (e.g. *dog*) had relatively low classification accuracy and were more likely to be confused with other similar categories (e.g., other animals) (see Figure 2B).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.015	0.199	-5.098	0.000
Age (in years)	0.238	0.020	11.905	0.000
Est. drawing frequency	-0.132	0.199	-0.662	0.508
Age*Drawing frequency	0.002	0.017	0.117	0.907
Tracing score	0.262	0.020	12.868	0.000
Time spent drawing	0.068	0.021	3.219	0.001
'Ink' used (mean intensity)	-0.071	0.021	-3.412	0.001
Number of strokes	0.016	0.018	0.886	0.376

Table 1

Model coefficients of a GLMM predicting the recognizability of each drawing (i.e. binary classification scores), including random intercepts for each category and participant.

Recognizable drawings contain more diagnostic features across development. The above results could reflect gradual improvements to children's ability to include diagnostic features of each category in their drawings. However, they are also consistent with an alternative account where younger children have similar overall competence in producing recognizable drawings when they are engaged with the task, but are less likely to stay on-task than older children and thus more often produce unrecognizable drawings. To tease these two possibilities apart, we compared how much diagnostic visual information was contained in drawings that were *correctly classified*. For example, among drawings that were correctly recognized as *clocks*, did older children also include features that more clearly set them apart from other similar

categories – for example, *watches*? Insofar as age-related improvements in classification accuracy primarily reflect a decrease in the proportion of unrecognizable drawings – rather than an increase in the quality of their recognizable drawings – we should expect younger and older children's recognizable drawings to contain similar amounts of diagnostic visual information.

Even for drawings that were correctly classified (33% of the balanced subset of drawings, N=7,468) there was still a reliable increase in the amount of diagnostic information they contained across age, as measured by the log-odds probability assigned by the logistic-regression classifier to the target category (see *Methods*, $B = 0.102$, $SE = 0.016$, $df = 3297.23$, $t = 6.524$, $P < 0.001$). Thus, this secondary analysis provides additional support for the account that the age-related improvements reflect gradual changes in children's ability to include diagnostic visual features in their drawings (see Figure 2B).

Rich information in misclassified drawings. Children's misclassified drawings were not merely incoherent scribbles; they still contained valuable visual information about the category they were intending to draw. Figure 3A shows the classifier probabilities assigned to all categories for drawings that were incorrectly recognized at the basic-level. These systematic model confusions highlight the structure in children's unrecognizable drawings: for example, drawings of *an octopus* were often misclassified as *a spider*, and drawings of *a book* were often confused with drawings of *a TV*.

Indeed, prior work has found that simple visual features can convey semantic information. For example, adults can identify the animacy and real-world object size of unrecognizable images by inferring that roundish things are more likely to be animals and less likely to be large, inanimate objects (e.g., buildings) (Long, Konkle, Cohen, & Alvarez, 2016; Long, Störmer, & Alvarez, 2017; Long, Yu, & Konkle, 2018). We thus also examined whether misclassified drawings contained information about the animacy and real-world size of the objects children were intending to draw. To do so, we analyzed the misclassifications made by the logistic regressions and found that these

misclassifications were highly informative with respect to the animacy of the intended category for children of all ages (see Figure 3B). Similarly, we found that misclassified drawings still contained information about the real-world size of the inanimate objects children were trying to depict (see Figure 3C). Even when children were unable to convey the basic-level category that they were intending to draw, their drawings still contained rich information about the visual features of that category.

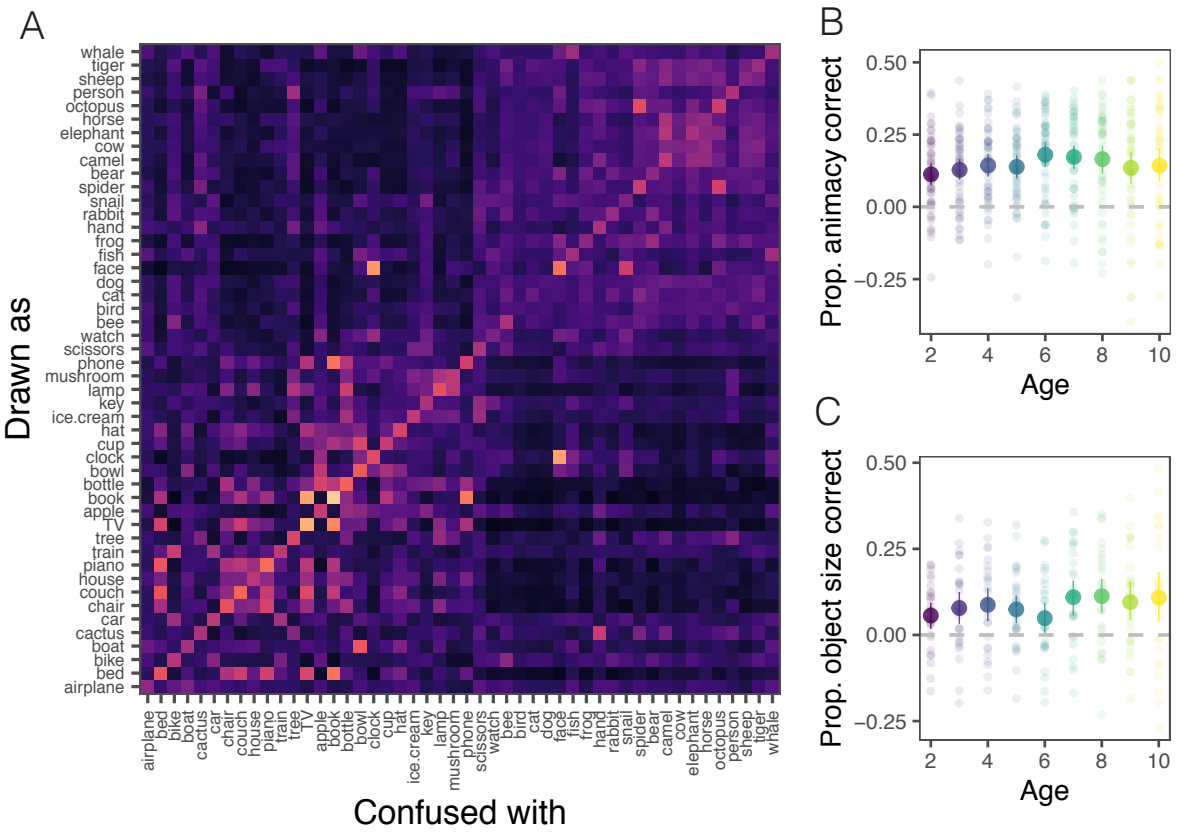

Figure 3. (A). Classifier probabilities for the subset of drawings that were misclassified. The y-axis shows the category children were intending to draw; the x-axis shows all of the categories in the dataset. Lighter values represent greater classifier probabilities assigned to a given category (B,C). Proportion of misclassified drawings that contained the correct animacy/object size information of the target category (relative to baseline). Each dot represents the proportion of drawings in a given category that had correct animacy/real-world size information relative to baseline at each age, respectively. Error bars represent bootstrapped 95% confidence intervals.

Relationship to visuomotor control. Finally, we anticipated that the recognizability of children's drawings would vary with their ability to control and plan their motor movements. Children spend countless hours across childhood both learning to write and practicing how to produce different shapes. Further, children's engagement with this drawing task could also reasonably vary as a function of age, with more skilled children spending more time, ink, or strokes on their drawings. We thus measured the amount of time and effort children put into their drawings, and we estimated children's visuomotor control via the simple shape tracing assessment task at the drawing kiosk. Children were instructed to trace both a relatively easy shape (a square) as well as a complex, novel shape that contained both curved and sharp segments (see Figure 1). For each participant, we used their performance on these two tracing trials to derive estimates of their tracing ability. Specifically, we obtained ratings of tracing accuracy from independent adult judges for a subset of tracings and then used these ratings to adapt an image registration algorithm (Sandkühler, Jud, Andermatt, & Cattin, 2018) to predict tracing scores for held-out tracings produced by children (see *Methods*). We found that tracing scores produced by the same participant were moderately correlated (r=.60, p<.001, N=6,746), despite the irregular shape being notably harder to trace than the square. Thus, despite the brevity of this tracing assessment, the resulting measure had moderate reliability.

Individuals' tracing abilities were good predictors of the recognizability of the drawings they produced. However, we still observed a robust main effect of age even when accounting for tracing abilities and other effort covariates (see Appendix Figure B2, Table 1), including the amount of time children spent drawing, the number of strokes in their drawings, and the amount of 'ink' that they used. Children's ability to control and plan their motor movements has a substantial effect on their ability to produce recognizable drawings—but does not entirely account for the observed developmental changes.

Development of drawing recognition

Why might children include more diagnostic visual features in their drawings as they grow older? One explanation for these developmental changes is that children improve their ability to connect their internal and external representations of these visual concepts as they acquire more knowledge about the visual distinctions between different visual concepts. In other words, children might more clearly represent the visual features that best distinguish between *rabbits* versus *dogs*. This account predicts that older children should also be better able to exploit this diagnostic visual information in drawings to recognize their intended meaning.

To test this idea, we installed a "guessing game" in the same kiosk at the local science museum (see Figure 1, top right) where children guessed the category that an earlier child's drawing referred to. These drawings were randomly sampled from the larger drawing dataset and varied in the amount of diagnostic visual information they contained. This design choice allows us to examine how children's visual recognition abilities vary when drawings contain differing amounts of diagnostic visual features.

Our goal in designing the visual recognition task was for it to be challenging yet not demand that children track a large number of comparisons. At the beginning of each session, children completed four practice trials in which they were cued with a photograph and asked to "tap the [vehicle/animal/object] that goes with the picture," choosing from an array of four photographs of different object categories (see Figure 1). Children were then cued with drawings of these categories and responded using the same photograph buttons; photograph matching trials were also interspersed throughout the session as attention checks. We sequentially deployed four different versions of this task, including a different set of four perceptually similar categories in each (e.g., *hat, bottle, cup, lamp*). After exclusions, our dataset from this task included 1,789 children ages 3–10 years (see *Methods*).

Children became steadily better at identifying the category that a drawing referred to (see Figure 4A). In contrast, performance on photograph matching trials was relatively similar across ages. All included children scored >75% correct on photograph

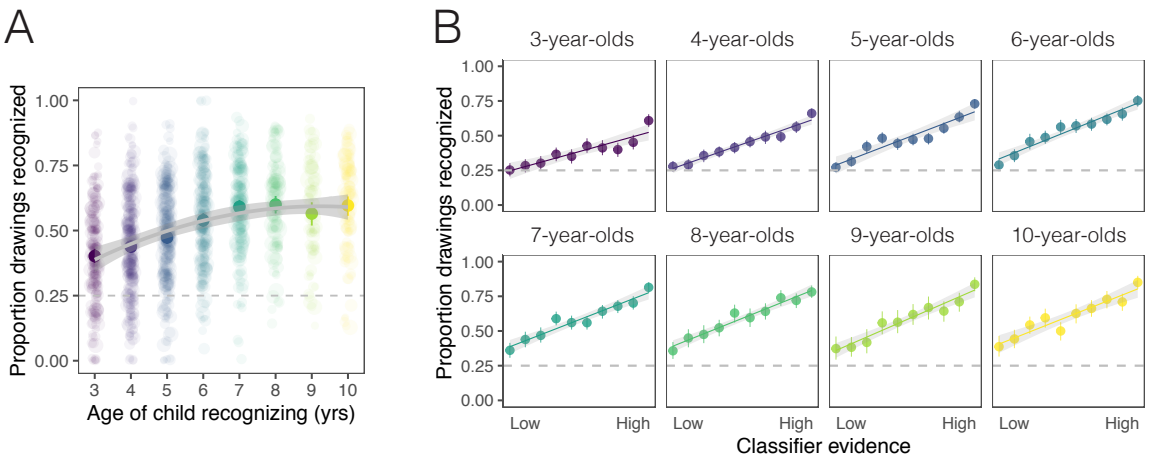

Figure 4. A. Drawing recognition as a function of the age of the child who participated in the guessing game; each dot represents data from one child who participated and is scaled by the number of trials they completed. B. Drawing recognition data plotted separately by the age of the child participating as a function of the amount of diagonstic visual features in each drawing, operationalized as the the *classifier evidence* assigned to each sketch relative to the distractor categories. Classifier evidence is binned into deciles for visualization purposes. Error bars represent bootstrapped 95% confidence intervals.

trials and average accuracy in each group ranged from M=90-93% correct. Thus, variation in drawing recognition accuracy is unlikely to be explained by generic differences in motivation or task engagement.

To test the hypothesis that children's drawing recognition ability reflects the amount of diagnostic information in a given drawing, we evaluated whether drawings with more diagnostic information were better recognized. For each drawing that appeared in the guessing game, we fit a 4-way logistic regression classifier trained on the drawings presented in each guessing game and measured diagnostic information as the log-odds ratio between the intended category and the foil categories. That is, the diagnostic information for a *dog* drawing was defined relative to its perceptual similarity to the other choices in the recognition task (i.e., *bird, fish, rabbit*). We then fit a generalized linear mixed effects predicting children's recognition performance with child's age, this classifier evidence measure, and their interaction as fixed effects (see *Methods*).

	Estimate	Std. Error	z value
(Intercept)	0.050	0.121	0.413
Classifier evidence	0.477	0.046	10.406
Recognizer Age	0.317	0.019	16.777
Classifier evidence*Recognizer Age	0.062	0.014	4.246

Table 2

Model coefficients of a GLMM predicting visual recognition performance as a function of recognizer age and classifier evidence.

Older children were more accurate at recognizing drawings and, across all ages, drawings with more diagnostic information were better recognized (see Table 2). Consistent with our hypothesis, older children were better able to capitalize on graded differences in the diagnostic visual information in drawings when recognizing them (see Figure 4B), evidenced by an interaction between classifier evidence and recognizer age. This result held when we restricted our analyses to a subset of children who performed at ceiling on photograph matching trials (see *Appendix* Table B1). Thus, these results support the hypothesis that children's ability to use diagnostic visual features during recognition changes across development.

General Discussion

To what extent do children's visual concepts change across childhood? To examine this question, we conducted a large-scale investigation of how children produce and recognize a wide range of visual concepts across development (2-10 years of age). We found robust improvements in children's ability to include diagnostic visual features in their drawings. We also found gradual changes in children's ability to capitalize on these same diagnostic features when recognizing drawings of visual concepts. Together, these results suggest parallel developmental changes in both visual production and recognition of drawings across childhood. Children may undergo a more protracted developmental trajectory for the development of visual concepts than previously

thought, in tandem with refinements in their perceptual abilities (Bova et al., 2007; Natu et al., 2016) and their semantic knowledge about object categories (Tversky, 1985; Vales, Stevens, & Fisher, 2020).

More broadly, the present work highlights how larger-scale datasets of naturalistic behaviors can contribute to theoretical debates in developmental science. By collecting rich data from many participants over a large developmental age range, we can more precisely estimate graded changes in children's abilities and the degree to which these trajectories vary between participants and across categories. Using this approach, we find evidence for continuous and variable changes in children's visual concepts across development – rather than a point at which children become "adult-like". We believe that this work paints a more accurate picture of developmental change and opens up new avenues for investigating the various factors that shape visual concepts throughout development.

Several learning mechanisms are consistent with the developmental changes we observed. One possibility is that children are becoming better visual communicators as they learn which features are most effective at conveying category membership through the process of producing drawings. In turn, this process of using drawings and other visual modalities to communicate various visual concepts may have downstream effects on children's ability to recognize drawings of them. Such a mechanism would be consistent with prior work suggesting that learning to produce letters by hand can support subsequent letter recognition (James, 2017; Longcamp, Zerbato-Poudou, & Velay, 2005), with recent findings pointing towards the variability of visual forms seen while learning to write as a key factor (Li & James, 2016). Training studies with adults also provide support for a link between visual production and recognition: practicing drawing categories that are perceptually similar to one another (i.e. *beds* vs *chairs*) can lead to refinements in the ability to distinguish between them (Fan et al., 2020) as well as to more discriminable representations in visual cortex (Fan et al., 2020). Thus, the process of iteratively producing and recognizing drawings of visual concepts could cause these parallel developmental changes in both domains. Contra a strong version of this

account, however, we did not find strong effects of drawing practice at the category-level in the present data: for example, camels were among the best recognized categories and estimated (by parents) to be among the least practiced by children.

A second possibility is that children are explicitly learning the diagnostic features of categories as they enrich their semantic knowledge. For example, children may learn that certain visual features are functional in nature: camels have humps to store water, clocks have numbers to tell time, and whales spout water because they need to breathe. As a result, children may come to more clearly represent which visual features are diagnostic of different categories and why. In turn, this semantic knowledge could be filtered into children's visual concepts and thus accessed both when children draw an object and when they recognize it. This possibility aligns with a wealth of evidence suggesting that continual learning about different categories throughout the early school years shapes children's categorization abilities. For example, children change in how they think about the diagnosticity of different semantic properties across development: in early childhood, the fastest cheetah – that is, the exemplar with the most extreme value on some property – tends to be seen as the best and the most representative cheetah (Foster-Hanson & Rhodes, 2019). At the same time, taxonomic groupings become increasingly important both in children's explicit conceptual judgements (Tversky, 1985) and when children spontaneously arrange different visual concepts (e.g., wild vs. farm animals, Vales et al., 2020). Thus, children's evolving semantic knowledge could shape the visual features children use both when producing and recognizing different visual concepts.

A third possibility, not mutually exclusive with the other two, is that children are implicitly learning the diagnostic features of these visual concepts through the process of visual categorization itself: through repetitively viewing and categorizing depictions, real-life examples, and photographs of these different categories. Indeed, the deep neural network used here to categorize drawings was trained solely on photographs of object categories – it has never seen a drawing, had visuomotor experience drawing, or learned the semantic properties of these categories. Thus in principle it is possible that

children could be learning the diagnostic features of these visual concepts without substantial involvement from other cognitive or sensorimotor systems.

There are various limitations to the generalizability of these findings that future work could address. First, while these datasets are large and sample heterogeneous populations, all drawings and recognition behaviors were collected at a single geographical location, limiting the generalizability of these results to children from other cultural or socioeconomic backgrounds. Second, while we imposed strong filtering requirements, we were not present while the children were drawing or guessing and thus cannot be sure that we eliminated all sources of noise or interference. Many sources of additional interference would only generate noise in our data, though, rather than creating specific age-related trends. Third, since these datasets are cross-sectional, they cannot address whether changes in visual production precede changes in visual recognition or vice versa.

Ultimately, these results call for systematic, experimental investigations into the kinds of experience – including visuomotor practice, semantic enrichment, and visual exposure – that may influence visual production and recognition in children. We propose that a full understanding of how children produce and recognize drawings of common object categories 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.

Methods & Materials

Drawing Station Details

While the interface was designed to be navigable by children, the first page of the drawing station showed a short consent form and asked parents to enter their child's age in years. Afterwards, video prompts of an experimenter guided the child through the rest of the experiment; an initial video stated that this game was "only for one person at a time" and asked children to "draw by themselves." Every session at the drawing station started with tracing trials before moving on to the category prompts ("What

about a [*couch*]? Can you draw a [*couch*]?"). Children could stop the experiment at any time by pressing a stop button; each trial ended after 30 seconds or after the child pressed the "next" button. Six different sets of eight category prompts rotated at the station, yielding drawings from a total of 48 categories (*Appendix* Figure B1; airplane, apple, bear, bed, bee, bike, bird, boat, book, bottle, bowl, cactus, camel, car, cat, chair, clock, couch, cow, cup, dog, elephant, face, fish, frog, hand, hat, horse, house, ice cream, key, lamp, mushroom, octopus, person, phone, piano, rabbit, scissors, sheep, snail, spider, tiger, train, tree, TV, watch, whale).

Drawing 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 subsequent 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 11797 subsequent sessions at the station, 3094 filled out the survey, and 719 reported interference, 6.09% of participants; these participants' drawings were not rendered or included in analysis.

Raw drawing data of object categories 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). A first subset of drawings (N = 15594 drawings) was filtered manually by one of the authors, resulting in N = 13205 drawings after exclusions (15.3% exclusion rate); subsequently, drawing filtering was crowd sourced via Prolific. 390 participants first completed a practice round

demonstrating valid and invalid drawings and then viewed 24 drawings from a intended category at a time and selected the invalid drawings they judged to come from from off-task participants. Participants were reminded that unrecognizable drawings were still "valid" drawings, and could proceed to the next category only after selecting a 'catch' invalid drawing. Each drawing in the dataset was viewed at least twice by two different participants. To be conservative, any drawing that was marked as 'invalid' by a participant was excluded from the dataset. These stringent filtering criteria resulted in the exclusion of an additional 9897 drawings, leading to an overall exclusion rate of 24.57% of the drawings and a final set of 37770 drawings from 8084 sessions. In the final dataset, there were more younger than older children, despite filtering; see *Appendix* Table A1 for a complete summary.

Measuring Tracing Accuracy

We developed an automated procedure for evaluating how accurately participants performed the tracing task that was 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.

To compute these error components, we applied an image registration algorithm, AirLab Sandkühler et al. (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 dataset (N=14372 tracings from 7612 children who completed at least one tracing trial). We averaged scores for both shapes within session to yield a single tracing score for each participant.

Measuring effort covariates

For each drawing trial, children had up to 30 seconds to complete their drawings with their fingers. We recorded both the final drawings and the parameters of each stroke produced by children at the drawing station, allowing us to estimate the amount of time children put into their drawings. As a second measure of effort, we also counted the number of strokes that children put into a given drawing. Finally, we estimated the proportion of the drawing canvas that was filled (e.g., 'ink used') by computing the proportion of each final drawing that contained non-white pixels.

Estimating drawing recognizability

Visual Encoder. To encode the high-level visual features of each sketch, we used the VGG-19 architecture Simonyan and 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 is the first fully connected layer of the network (FC6), as prior work suggests that it contain more explicit representations of object identity than earlier layers (Fan et al., 2018; Long, Fan, & Frank, 2018; Yamins et al., 2014). Raw feature representations in this layer consist of flat 4096-dimensional vectors, to which we applied channel-wise normalization across all filtered drawings in the dataset.

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 48 categories (N=22,272 drawings total). We then trained a 48-way logistic classifier with L2 regularization (tolerance = .1, regularization = .1), and used this classifier to estimate the category labels for a random held-out subset of 96 drawings (2 drawings from each category). No additional metadata about the age of the child who produced each sketch was provided to the decoder. This procedure was repeated for entire dataset (K=232 fold) yielding both a binary a recognition score and the softmax probability assigned to each target class in the dataset. We define *classifier evidence* as the log-odds ratio of the probability assigned to the target category vs. the other categories in the dataset; this metric thus captures the degree to which a given drawing contains features that are diagnostic of the target category (and not of the other categories in the dataset); these log-transformed values are also more suitable for the linear-mixed effects models used in analyses.

Mixed-effect models. Two mixed effects models were fit to assess the degree to which children produced more recognizable drawings across childhood. A first generalized mixed effect model was fit to the binary classification scores for each drawing using a logit linking function. A second linear mixed effect model was fit to the log-odds target probability assigned to each drawing, restricting our analyses to correctly classified drawings. In both cases, we included fixed effects of children's age

(in years), estimated drawing frequency for each category (via parental report), their interaction, children's estimated tracing score (see above), the time children spent drawing (in seconds), the mean intensity of the drawing (i.e. percentage of non-white pixels), and the number of strokes children used. All predictors were scaled to have a mean of 0 and a standard deviation of 1. Random intercepts were included for each participant and each category.

Animacy & object size information in misclassified drawings. For each drawing, we calculated whether the category assigned by the logistic regression classifier was of the same animacy as the target category, assigning a binary animacy classification score for each drawing. The same procedure was repeated for inanimate objects with respect to their real-world object size (big objects: larger than a chair, small objects: can be held with one hand, see (Konkle & Oliva, 2011; Long et al., 2016)). These binary scores were averaged for each age and category, yielding a value between 0 and 1 representing the proportion of the drawings that were identified as having the correct animacy/size. As the proportion of animals/inanimate objects and big/small inanimate objects was not exactly balanced in the dataset, we subtracted the baseline prevalence for each broad category (i.e for animals, objects, big objects, and small objects) from this proportion. These values are plotted in Figures 3B,C, as are the bootstrapped 95% confidence intervals calculated using the baseline-corrected category values.

Visual recognition task

Behavioral task. On each trial of the guessing game, a photograph or drawing of an object category was presented on the screen, and children were asked to "tap the [animal/vehicle/object] that goes the with the [drawing/picture]"; response choices were indicated by circular buttons that were filled photographs of canonical exemplars from each category, as well as the name of the category written above; the position of these response buttons was randomized for each participant. A fifth response choice was a button with a question-mark icon that could be used by participants to indicate they

didn't know which category the drawing belonged to. To familiarize participants with the interface, the first four trials of every game were four photograph trials, one for each of the response choices. To encourage accurate guessing, a pleasant sound was played when the correct category was chosen, and the box surrounding the image briefly turned green; no feedback was given for incorrect trials. Every ten trials, a catch trial appeared where participants were required to match a very similar photograph to the photographic response buttons.

Drawing selection. We selected four subsets of categories for the guessing game at the station: small animals (*dog, fish, rabbit, bird*), vehicles (*train, car, airplane, boat*), small, inanimate objects (*hat, bottle, cup, lamp*), and large animals (*camel, sheep, bear, tiger*). Each version of the guessing game ran separately for approximately two months. For each game, we randomly selected drawings (20-25 per category, depending on availability) made by children ages 4-9 at the drawing station. We chose this age range to cover a wide range of drawing abilities and to ensure equal numbers of drawings were included per age group (as 9-10 year-old's are infrequent visitors to the museum). This resulted in 516—616 drawings for each guessing game from which 48 drawings were randomly sampled for each participant (8 drawings made by 4-,5-,6-,7-,8-, and 9-year-olds). If children completed the entire session, this resulted in a total of 56 trials for each participant (48 drawing trials and 8 photograph matching trials).

Recognition data inclusion. As with the drawing data, we excluded any sessions where there was reported interference from parents or other children. As 2-year-old's showed significantly better performance than 3-year-old in our first two guessing games–signaling some interference from their caregivers or siblings that was not reported in the surveys–we chose to exclude 2-year-old's from subsequent analyses. We excluded children who started the game but did not complete more than 1 trial after the practice trials ($N = 1068$ participants) and the 238 adults who participated. We also excluded all trials with reaction times slower than 10s or faster than 100ms, judging these to be off-task responses. Next, we excluded participants on the basis of their performance on practice and catch photograph matching trials. Given that these

catch trials presented a very easy recognition task, we excluded participants who did not achieve at least 75% accuracy on these trials ($N = 795$). The remaining 1789 participants who met this criterion completed an average of M=21.69 trials. On total, we analyzed 36,615 trials where children recognized each other's drawings. These analysis choices were pre-registered after examining data from two of the guessing games and then applied to the entire dataset (see https://osf.io/dahkm).

Recognition data analyses. To calculate the classifier evidence associated with each sketch that children recognized, we used the same visual encoder to extract visual features for each sketch (see *Visual Encoder*), and iteratively trained logistic regression classifiers (see *Logistic Regression Classifier*). For these analyses, we restricted the classification set to the drawings that were presented in each version of the guessing game to match the task conditions of the guessing game. We trained a separate logistic regression for each sketch that was presented using leave-one-out cross-validation. This procedure thus yielded probabilities assigned to each of four categories in each guessing game; these probabilities were used to calculate the log-odds ratios for the target category of each sketch which we refer to as *classifier evidence*. Due to random sampling, not every sketch included in the game had valid guesses associated with it; these sketches were thus not included in analyses. We then modeled children's recognition behavior in a generalized linear mixed-effect model, where recognizer age (in years), classifier evidence, and their interaction were specified as fixed effects. All predictors were scaled between 0 and 1. We included random intercepts for the intended category of the sketch and for each subject who participated in the guessing game; random slopes were also included for the effect of classifier evidence on each intended category.

Acknowledgements

We gratefully acknowledge the San Jose Children's Discovery Museum for their collaboration and for hosting the drawing station where these data were collected. We also thank the 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

Allen Preissler, M., & Carey, S. (2004). Do both pictures and words function as symbols for 18-and 24-month-old children? *Journal of Cognition and Development*, *5*(2), 185–212.

Balas, B., & Saville, A. (2020). Neural sensitivity to natural image statistics changes during middle childhood. *Developmental Psychobiology*.

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.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *ERIC*.

Cohen, M. A., Dilks, D. D., Koldewyn, K., Weigelt, S., Feather, J., Kell, A. J., ... others (2019). Representational similarity precedes category selectivity in the developing ventral visual pathway. *NeuroImage*, *197*, 565–574.

Cohn, N. (2012). Explaining 'i can't draw': Parallels between the structure and development of language and drawing. *Human Development*, *55*(4), 167–192.

Dekker, T., Mareschal, D., Sereno, M. I., & Johnson, M. H. (2011). Dorsal and ventral stream activation and object recognition performance in school-age children. *NeuroImage*, *57*(3), 659–670.

DeLoache, J. S., Strauss, M. S., & Maynard, J. (1979). Picture perception in infancy. *Infant behavior and development*, *2*, 77–89.

Fan, J. E., Wammes, J. D., Gunn, J. B., Yamins, D. L., Norman, K. A., & Turk-Browne, N. B. (2020). Relating visual production and recognition of objects in human visual cortex. *Journal of Neuroscience*, *40*(8), 1710–1721.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object

representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Foster-Hanson, E., & Rhodes, M. (2019). Is the most representative skunk the average or the stinkiest? developmental changes in representations of biological categories. *Cognitive psychology*, *110*, 1–15.

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.

Gomez, J., Natu, V., Jeska, B., Barnett, M., & Grill-Spector, K. (2018). Development differentially sculpts receptive fields across early and high-level human visual cortex. *Nature communications*, *9*(1), 788.

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

Gregory, R. L. (1973). *Eye and brain: The psychology of seeing*. McGraw-Hill.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

James, K. H. (2017). The importance of handwriting experience on the development of the literate brain. *Current Directions in Psychological Science*, *26*(6), 502–508.

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.

Kersey, A. J., Clark, T. S., Lussier, C. A., Mahon, B. Z., & Cantlon, J. F. (2015). Development of tool representations in the dorsal and ventral visual object processing pathways. *Cerebral Cortex*, *26*(7), 3135–3145.

Konkle, T., & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of experimental psychology: human perception and performance*, *37*(1), 23.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Li, J. X., & James, K. H. (2016). Handwriting generates variable visual output to facilitate symbol learning. *Journal of Experimental Psychology: General*, *145*(3), 298.

Light, P., & Simmons, B. (1983). The effects of a communication task upon the representation of depth relationships in young children's drawings. *Journal of Experimental Child Psychology*, *35*(1), 81–92.

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.*

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95.

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision*, *17*(6), 20–20.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024.

Longcamp, M., Zerbato-Poudou, M.-T., & Velay, J.-L. (2005). The influence of writing practice on letter recognition in preschool children: A comparison between handwriting and typing. *Acta psychologica*, *119*(1), 67–79.

Mash, C. (2006). Multidimensional shape similarity in the development of visual object

classification. *Journal of Experimental Child Psychology*, *95*(2), 128–152.

Natu, V. S., Barnett, M. A., Hartley, J., Gomez, J., Stigliani, A., & Grill-Spector, K. (2016). Development of neural sensitivity to face identity correlates with perceptual discriminability. *Journal of Neuroscience*, *36*(42), 10893–10907.

Nishimura, M., Scherf, K. S., Zachariou, V., Tarr, M. J., & Behrmann, M. (2015). Size precedes view: developmental emergence of invariant object representations in lateral occipital complex. *Journal of cognitive neuroscience*, *27*(3), 474–491.

Nishimura, M., Scherf, S., & Behrmann, M. (2009). Development of object recognition in humans. *F1000 biology reports*, *1*.

Pereira, A. F., & Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental science*, *12*(1), 67–80.

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

Quinn, P. C., & Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *Journal of experimental child psychology*, *63*(1), 189–211.

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.

Sandkühler, R., Jud, C., Andermatt, S., & Cattin, P. C. (2018). Airlab: Autograd image registration laboratory. *arXiv preprint arXiv:1806.09907*.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and brain Sciences*, *21*(1), 1–17.

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

Sitton, R., & Light, P. (1992). Drawing to differentiate: Flexibility in young children's human figure drawings. *British Journal of Developmental Psychology*, *10*(1), 25–33.

Sloutsky, V. M., & Fisher, A. V. (2011). The development of categorization. In

Psychology of learning and motivation (Vol. 54, pp. 141–166). Elsevier.

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition*, *38*(2), 179–211.

Tversky, B. (1985). Development of taxonomic organization of named and pictured categories. *Developmental Psychology*, *21*(6), 1111.

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.

Weigelt, S., Koldewyn, K., Dilks, D. D., Balas, B., McKone, E., & Kanwisher, N. (2014). Domain-specific development of face memory but not face perception. *Developmental Science*, *17*(1), 47–58.

Willats, J. (2006). *Making sense of children's drawings*. Psychology Press.

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.

Appendix A

Demographics of drawing participants

Age	Number of participants	Number of drawings
2-year-olds	1231	3651
3-year-olds	1402	5342
4-year-olds	1451	6559
5-year-olds	1189	6411
6-year-olds	878	4990
7-year-olds	660	3817
8-year-olds	478	2570
9-year-olds	309	1800
10+-year-olds	486	2630

Table A1

Number of participants and drawings included in the filtered dataset by each age group.

Appendix B

Supplemental analyses & figures

Drawing recognition: Including only high-performing children.

To ensure that these results were not driven by differences in motivation or general task performance, we also conducted our main analyses on a very restricted subset of our participants. We excluded any participant that did not achieve 100% on the photograph matching trials or that scored less than 50% on the drawing recognition trials. While this excluded nearly two-thirds of our participants, there were nonetheless N=649 participants in this subset. Nonetheless, we still found the same pattern of results (see Table B1): older children were still better at recognizing drawings and at using diagnostic visual information in these drawings when recognizing them.

	Estimate	Std. Error	z value
(Intercept)	0.668	0.099	6.717
Classifier evidence (scaled)	0.518	0.051	10.059
Recognizer age (scaled)	0.141	0.023	6.190
Classifier evidence*Recognizer Age	0.056	0.023	2.464

Table B1

Model coefficients of a GLMM predicting visual recognition performance, excluding any participant who missed even one of the photograph trials or who scored less than 50% on drawing recognition trials.

Additional figures

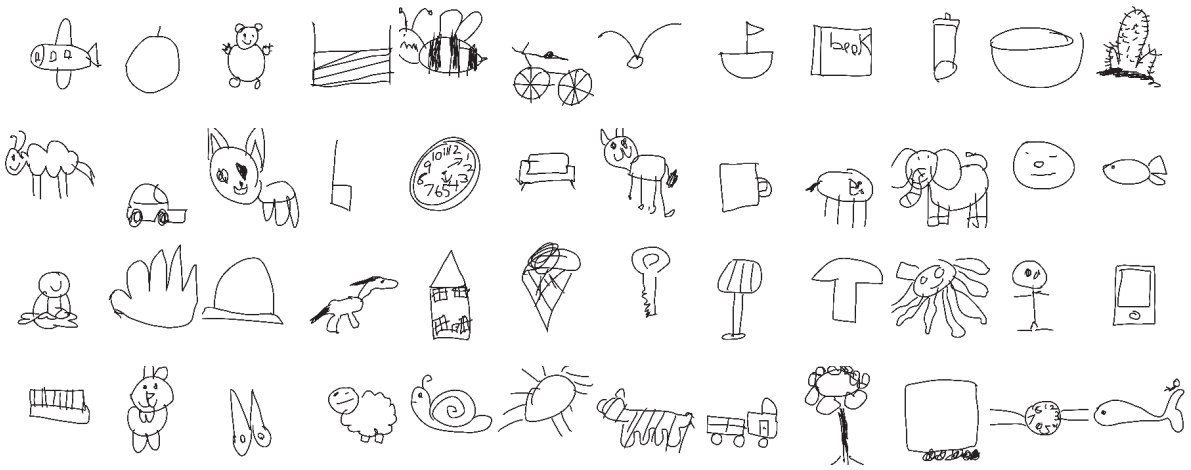

Figure B1. Examples of correctly classified drawings from each of the 48 categories presented at the experiment station in alphabetical order: airplane, apple, bear, bed, bee, bike, bird, boat, book, bottle, bowl, cactus, (2nd row): camel, car, cat, chair, clock, couch, cow, cup, dog, elephant, face, fish, (3rd row): frog, hand, hat, horse, house, ice cream, key, lamp, mushroom, octopus, person, phone, (4th row): piano, rabbit, scissors, sheep, snail, spider, tiger, train, tree, TV, watch, whale.

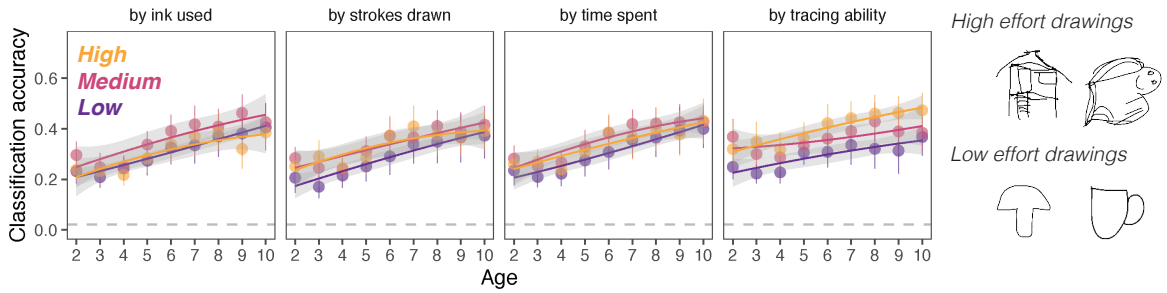

Figure B2. (Left): Classification accuracy by age, split into bins according to whether children expended a greater/lesser amount of strokes, ink, or time, and by their estimated tracing abilities (see Methods). (Right): Example drawings where children spent higher/lower amounts of *effort*—greater/lower than average number of strokes, time spent drawing, or 'ink' used.