Parallel developmental changes in children's production and recognition of line drawings of visual concepts

Bria Long[1], Judith E. Fan[1,2], Holly Huey[2], Zixian Chai[1], Michael C. Frank[1]
[1]Department of Psychology, Stanford University
[2]Department of Psychology, University of California San Diego

Abstract

Childhood is marked by the rapid accumulation of knowledge and the prolific production of drawings. We conducted a systematic study of how children create and recognize line drawings of visual concepts. We recruited 2-10-year-olds to draw 48 categories via a kiosk, resulting in >37K drawings. We analyzed changes in the category-diagnostic information in these drawings using vision algorithms and annotations of object parts. We found developmental gains in children's inclusion of category-diagnostic information that was not reducible to variation in visuomotor control or effort. Moreover, unrecognizable drawings contained information about the animacy and size of the category children tried to draw. Using "guessing games" at the same kiosk, we found that children improved across childhood at recognizing each other's line drawings. This work leverages vision algorithms to characterize developmental changes in a large dataset of children's drawings and suggests that changes in children's drawings reflect refinements in internal representations.

Parallel developmental changes in children's production and recognition of line drawings of visual concepts

Introduction

What makes a drawing of a *rabbit* look like a *rabbit* and not a *dog*? As adults, our *visual concepts* – our sense of what particular objects look like – are seamlessly integrated into our visual experience. With a single glance, incoming patterns of light make contact with our visual concepts, supporting the rapid categorization of a wide variety of inputs, from real-life exemplars to sparse line drawings and stylized animations (Biederman & Ju, 1988; Gibson, 1971; Hertzmann, 2020; Sayim, 2011). We can also access our visual concepts in the absence of perceptual input – going beyond what we have previously experienced to imagine new visual entities and create external representations of them (Clottes, 2008; Finke & Slayton, 1988; Gregory, 1973).

Yet while these feats of perceiving and creating can feel effortless, the representations that support them are acquired and refined gradually as children learn about the visual world (Rosch, 1978). Children begin building visual concepts in earnest during the second year of life as they learn which labels refer to both depictions and real-life exemplars of categories (DeLoache, Pierroutsakos, & Uttal, 2003). And by their second birthday, children can learn category labels for novel objects after exposure to just one or a few exemplars (Carey & Bartlett, 1978; Pereira & Smith, 2009; Soja, Carey, & Spelke, 1991) and succeed even for sparse 3D representations of these objects devoid of color and texture-based cues (Pereira & Smith, 2009).

But children take many years to learn how to appropriately generalize and discriminate between visual concepts. For example, children gradually improve in their ability to accurately group together categories based on taxonomy versus salient perceptual features (e.g., grouping a *snake* with a *lizard* vs. a *hose*) (Fisher, Godwin, & Matlen, 2015; Tversky, 1985). Further, children's visual recognition abilities also have a protracted developmental trajectory throughout middle childhood (Bova et al., 2007; Juttner, Wakui, Petters, & Davidoff, 2016; Nishimura, Scherf, & Behrmann, 2009) as children become steadily better at discriminating between similar exemplars of scenes, objects, bodies, and faces between 5-10 years of age (Weigelt et al., 2014), and increasingly skilled at recognizing objects across unusual poses or 3D rotations, reaching adult-level performance only in adolescence (Bova et al., 2007; Dekker, Mareschal, Sereno, & Johnson, 2011; Nishimura, Scherf, Zachariou, Tarr, & Behrmann, 2015). In turn, changes in children's recognition abilities are related to changes in how the 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 similar faces is correlated with the sensitivity of face-selective regions to these particular faces (Natu et al., 2016). These changes in children's ability to discriminate exemplars may be driven by children's increasing attention to the relationships between object parts and their overall configuration (Juttner, Muller, & Rentschler, 2006; Juttner et al., 2016; Mash, 2006). Together, these findings suggest that visual concepts are refined throughout childhood as children's perceptual abilities mature and children learn how to discriminate between similar categories.

Psychologists have typically probed children's visual concepts by asking children to make discrete choices between small samples of stimuli that vary along dimensions chosen by an experimenter. While valuable for testing specific hypotheses, this strategy is also characterized by severe limits on the amount of information that can be acquired on any given experimental trial. By contrast, generative tasks such as drawing production can overcome these limits by enabling the collection of more information about the contents of children's visual concepts on every trial. Such tasks are feasible to administer in experimental settings given that almost all children prolifically produce drawings of visual concepts from an early age (Piaget, 1929). And there is substantial precedent for examination of children's drawings to probe their knowledge about the visual world (Fury, Carlson, & Sroufe, 1997; Karmiloff-Smith, 1990; Kellogg, 1969; Piaget, 1929). Freehand drawing production tasks thus provide a valuable tool for characterizing developmental changes in visual concepts. Here we create a large digital dataset of children's drawings and leverage innovations in machine learning to characterize how changes in children's drawings are related to their growing understanding of various visual concepts.

Drawings as a window into visual representations

Our work builds on a long literature that has argued that children's drawings of objects reflect not only what they can directly observe, but what they *know* about these objects (see "intellectual realism" in Freeman & Janikoun, 1972; Luquet, 1927). For example, even when drawing from observation, children tend to include features that are not visible from their vantage point but are nevertheless diagnostic of category membership (e.g., an occluded handle on a *mug*) (Barrett & Light, 1976; Bremner & Moore, 1984). Further, direct visual or haptic experience with novel objects tends to change what information children draw (Bremner & Moore, 1984). These initial studies have focused on a small number of visual concepts – especially the human figure (Goodenough, 1963) – finding that younger children (4-5 years) tend to include fewer category-diagnostic cues in their drawings, such as those distinguishing an "adult" from a "child," than somewhat older children (6 years), who tend to enrich their drawings with more diagnostic part information (Cox & Ralph, 1996; Sitton & Light, 1992). However, the generality of the conclusions based on this work has been unclear given the narrow range of concepts tested and the lack of generic methods for measuring diagnostic information in drawings. Further, little work has systematically related children's ability to include diagnostic visual information in drawings to their emerging abilities to control and plan their motor movements – which certainly influence how and what children draw (Freeman, 1987; Rehrig & Stromswold, 2018).

Yet research in adults does suggest that what we draw is tightly linked to what we know about objects and how we perceive them. For example, patients with semantic dementia tend to produce drawings without distinctive visual features (Bozeat et al., 2003) or include erroneous features (e.g., a duck with four legs). One recent study found that adults can produce detailed drawings of scenes after only viewing them for a few seconds, interleaved among other scenes (Bainbridge, Hall, & Baker, 2019). Another study found that recognizing an object and producing a drawing of an object recruit a shared neural representation in early visual cortex (Fan et al., 2020). Further, practice producing drawings of objects can impact perceptual judgments about them. In one study, adult participants who repeatedly drew similar objects (i.e. *beds* vs *chairs*) were better able to distinguish them in a categorization task (Fan, Yamins, & Turk-Browne, 2018). Drawing expertise is also associated with enhanced visual encoding of object parts and their relationships (Chamberlain et al., 2019; Perdreau & Cavanagh, 2013a, 2013b, 2014), but not differences in low-level visual processing (Chamberlain et al., 2019; Chamberlain, Kozbelt, Drake, & Wagemans, 2021; Kozbelt, 2001; Perdreau & Cavanagh, 2013b) or shape tracing skills (Tchalenko, 2009).

The current study

Building on these traditions, in the current paper we characterize developmental changes in how children produce and recognize line drawings as an additional lens into children's growing understanding of these visual concepts. We anticipated that children's ability to produce and recognize line drawings would continue to develop beyond the preschool and elementary school years (Dekker et al., 2011; Gomez et al., 2018; Weigelt et al., 2014) and that some – but not all – age-related variation in drawing ability would be due to improvements in planning and motor control (Freeman, 1987; Rehrig & Stromswold, 2018). In particular, as children learn the visual information most diagnostic of a visual concept (Rosch, 1978), this visual knowledge may manifest in both: (1) an enhanced ability to produce line drawings that contain category-diagnostic information and (2) a greater sensitivity to this same visual information when recognizing line drawings made by other children.

We thus first collected digital drawings of 48 different visual concepts from a large sample of children spanning a wide age range (2-10 years), resulting in a corpus containing >37K drawings. In quantify developmental changes in these drawings at scale, we leveraged techniques from modern machine learning and computer vision, in particular the latent feature representations learned by large neural networks trained on visual discrimination tasks (Radford et al., 2021; Simonyan & Zisserman, 2014), which have been shown in prior work to capture meaningful variation in human perceptual judgments about both natural images and drawings (Battleday, Peterson, & Griffiths, 2020; Fan et al., 2018). We use these latent feature representations both to quantify the category-diagnostic variation in each drawing and to analyze the similarity structure in children's unrecognizable drawings. In addition, we crowd-sourced part labels for each stroke in a subset of these drawings to quantify how the parts children included in their drawings changed across development. Finally, we administered drawing recognition tasks to measure how well children of different ages could identify which visual concept a given drawing was intended to convey.

This study makes a number of contributions relative to the prior literature. First, we collect, annotate, and share a large sample of children's drawings from scribbles through sophisticated sketches, creating valuable resources for future research. Second, we develop an analytic approach suitable for exploring these drawings, which yields a number of intriguing findings around drawing development – including the presence of semantic information even in children's unrecognizable drawings. Finally, we find evidence for the relation between developmental changes in children's drawing abilities and their growing understanding of the visual concepts they are drawing. Older children include more diagnostic visual information and relevant object parts when producing line drawings, and these gains are not easily explainable by category exposure frequency or visuomotor development. Further, children's developing ability to recognize drawings is related to the presence of category-diagnostic information in these drawings. Together, we provide a new set of tools and insights into the development of drawings and visual representations in childhood, which we hope will spur future research on this topic.

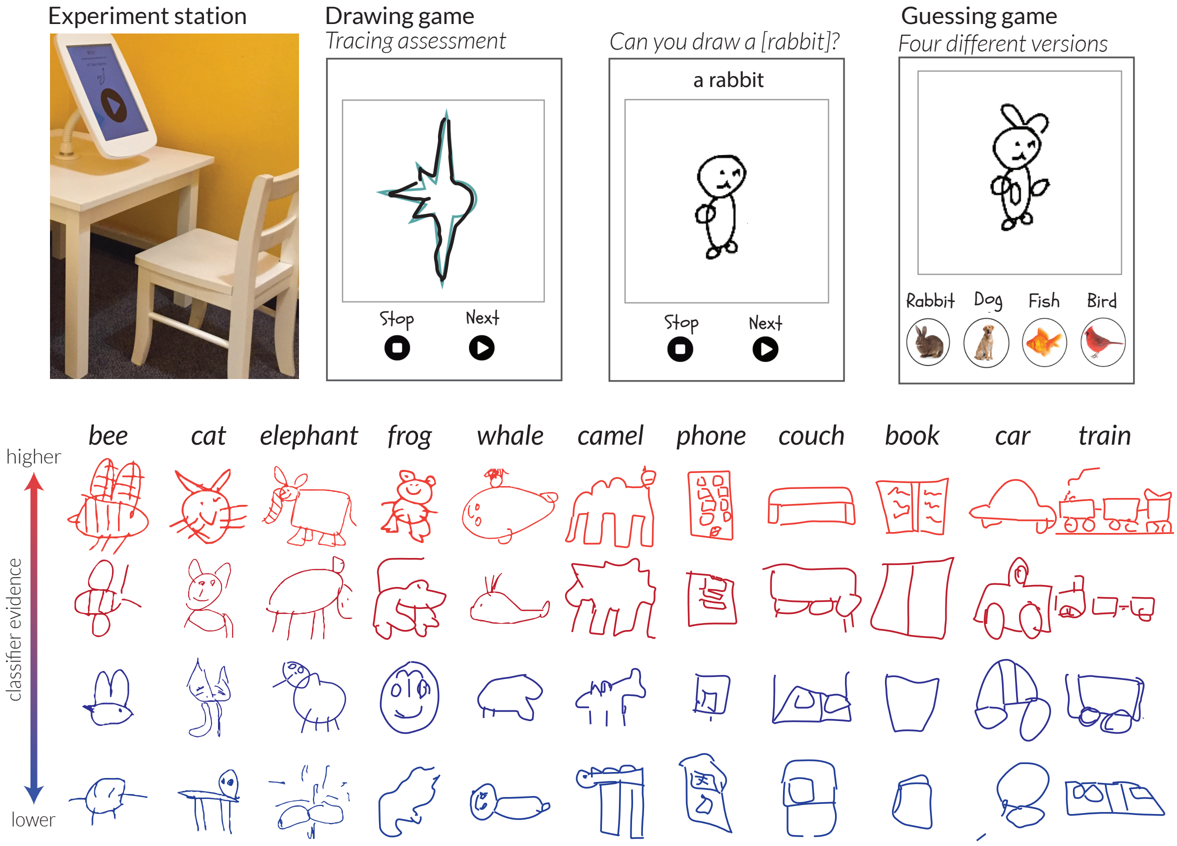

Figure 1. Top row: Museum kiosk where children participated, and examples of the tracing, drawing, and guessing trials. Bottom row: Example drawings from several categories; more red drawings contain more diagnostic visual features (as assessed by classifier evidence using VGG-19 FC6 features, see *Methods*).

Results

Development of drawing production

A free-standing, child-friendly kiosk was installed at a local children's science museum (see Figure 1, top row). Children used a touchscreen tablet attached to the kiosk to produce their drawings. To evaluate how 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 a set of shape-tracing trials in the drawing production task to measure children's tracing skills (see Figure 1, top row). After completing these tracing trials, children were verbally prompted to draw different visual concepts. These categories were selected to include both animals and inanimate objects, as well as categories that are either commonly drawn (e.g., *cup, face, cat*) or less commonly drawn (e.g., *octopus, piano, camel*) by children (see *Methods* for more details, see Appendix A6). The final, filtered dataset contained 37,770 drawings of 48 categories from N=8084 children (average age: 5.33 years old; range: 2–10 years old; see *Appendix* A for detailed age demographics).

Measuring category-diagnostic information in such a large dataset of children's drawings spanning a wide variety of concepts poses a major analytical challenge. Until recently, researchers attempting to analyze even small drawing datasets had to develop *ad hoc* criteria for scoring drawings based on their intuitions about what the distinctive visual features could be for each target concept (e.g., handles for *mugs*) (e.g. Barrett & Light, 1976; Goodenough, 1963). Fortunately, recent advances in computer vision have made it possible to measure category-diagnostic information in images at scale by leveraging latent feature representations learned by large neural networks (Radford et al., 2021; Simonyan & Zisserman, 2014), although at some cost to interpretability, as these learned features are not guaranteed to map onto nameable object parts (e.g., "handles"). Informed by this context, we use two approaches with complementary strengths to analyze our drawing dataset: first, we use model classifications to estimate the amount of category-diagnostic information in each drawing; second, we use crowdsourcing to identify which parts children included in their drawings in a general and scalable way.

Our first approach leverages the latent feature representations learned by large neural-network models to derive measures of a drawing's *recognizability* — how much category-diagnostic information it contains. Specifically, we analyzed the degree to which high-level visual features of each drawing could be used to decode the category that children intended to draw (e.g., *dog*), using features from VGG-19, a deep convolutional neural network pre-trained on Imagenet classification (Simonyan & Zisserman, 2014). Activations for each sketch were taken from the second-to-last layer of this model as prior work has shown that activations from deeper layers of convolutional neural networks with a similar architecture 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*), balanced across categories. 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.

We then validated these VGG-19 model classifications by using embeddings from a modern contrastive language-image pre-training model (CLIP, Radford et al. 2021), which jointly trains an image and a text encoder to predict image and text pairings. While less work has related the embeddings of this model class to either human behavioral or neural representations, CLIP outperforms other deep CNNs at recognizing visual concepts across different visual formats (Radford et al., 2021) and recent work suggests that CLIP embeddings show equal or better performance in predicting ventral stream responses (Conwell, Prince, Kay, Alvarez, & Konkle, 2022). CLIP classifications were obtained by assessing the similarity between model embeddings for each sketch to each category label, as in Radford et al., 2021. This method thus also yields both binary classification scores and probability scores for each of the 48 categories in the dataset.

As our second approach, we used a crowd-sourcing method to recruit human annotators to tag every stroke in a subset of N=2160 drawings with a part label (see *Methods*). To ensure that these drawings were representative of the larger dataset, we chose 16 visual concepts (half animate, half inanimate, see *Methods*) and randomly sampled drawings from children 4-8 years of age. Using these annotations, we then analyzed changes in which parts children drew and how much they emphasized those parts in their drawings. The goal of these additional analyses was to provide insight into which specific elements within children's drawings change across development, giving rise to any changes in category-diagnostic information measured using model classifications.

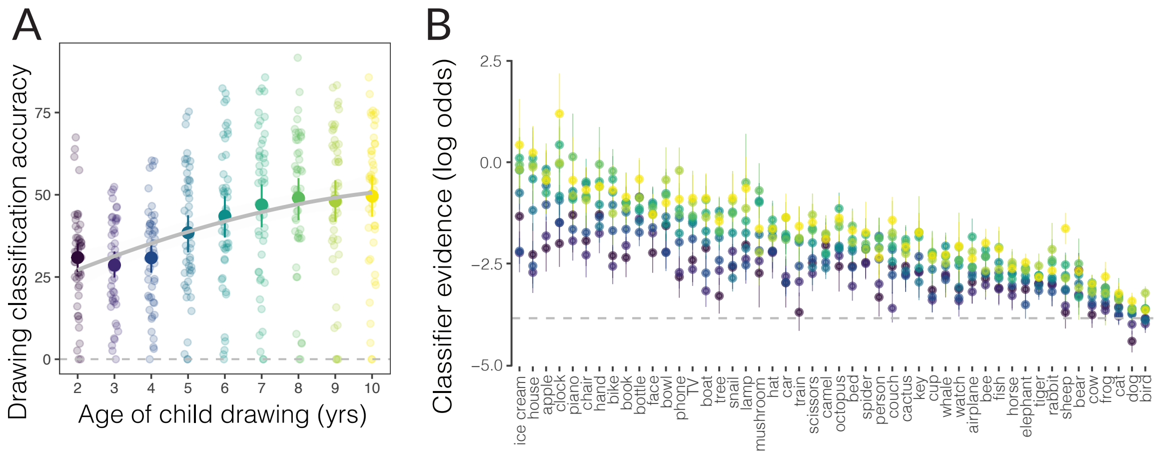

Figure 2. A. Proportion of drawings recognized as a function of children's age; each dot represents the proportion of drawings that were correctly classified in a given category; the grey chance line represents 1/48 (number of categories in the dataset). B: The y-axis represents the log-odds probabilities (i.e. classifier evidence), binned by the age of the child who produced the drawing. Categories on the x-axis are ordered by average log odds probabilities for each category in descending order. Error bars represent bootstrapped 95% confidence intervals in both plots.

Drawings of visual concepts become more recognizable across childhood. We found that children's drawings increased in recognizability steadily with age, as measured using model classification performance (VGG-19, Figure 2A, Table 1; see validation using CLIP in Appendix, Figure A2, Table A3).[1] In an additional study, we replicated this finding in a separate controlled experiment in which a researcher was present while children produced their drawings (N=121 children, ages 4-9 years), suggesting that the developmental changes we measured were not an artifact of data collection at the museum kiosk (this is a subset of the data published in BLINDED (in press), see Appendix, Figure A4, Table A4).

[1] We found that using features from deeper layers of VGG-19, rather than earlier layers, was critical to recovering these age-related changes, suggesting that drawings produced by older children primarily differed from those by younger children with respect to mid- and high-level visual features (see Appendix, Figure A5)

More frequently drawn categories are not more recognizable. What explains these gradual increases in recognizability? One way of accounting for these age-related differences is to suppose that younger children have had less practice drawing and are thus less well equipped to express what they know using this medium, despite having achieved a mature understanding of these visual categories. This account predicts that changes in recognizability are primarily driven by children's drawing experience with specific categories, either on their own or with caregivers and peers in educational contexts. If so, then frequently drawn categories (e.g., *trees*, *people*, *dogs*) should show the strongest developmental trends. To test this possibility, we asked parents to report how often their child produces drawings of each category (N=50 parents of children aged 3-10 years, *Methods*), revealing substantial variation in the frequency with which children tend to draw each of the categories in our stimulus set (see Appendix, Figure A6). We found converging evidence that drawings of more frequently practiced categories were no more recognizable and were not associated with 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-effects model (see Table 1). This result was robust to the choice of model (VGG-19 and CLIP, see Appendix A3) and held when using human recognition scores in a separate controlled experiment (see Appendix, Figure A4, Appendix Table A4). Instead, we saw that many infrequently drawn categories (e.g. *ice cream*) 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).

Figure 2B shows these developmental trends broken down by the category that children were intending to draw, highlighting the large amounts of variability across categories (see Appendix, A3 for validation using CLIP embeddings). We additionally examined whether other measures of frequency of experience in children's daily life might predict this item variation—–for example, frequency in child-directed speech or all English-language books (see Appendix A5). However, we again found no discernable relationship between these measures of frequency and the recognizability of children's drawings.

Visuomotor control explains some but not all of changes in drawing recognizability. 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 therefore measured the amount of time and effort children put into their drawings, and 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 ($r = .60$, $t = 61.93$, $df = 6754$, $p < .001$, $N = 6{,}756$), despite the irregular shape being harder to trace than the square. Thus, despite the brevity of this tracing assessment, the resulting measure had moderate reliability.

If age-related changes in drawing recognizability primarily reflect changes in visuomotor control (Freeman, 1987), then accounting for these more direct measures of visuomotor control ought to explain away the age-related variance we have observed so far. However, we still observed a robust main effect of age even after accounting for tracing abilities (Table 1) and other effort covariates (see 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 (see *Methods*); this effect was robust to model choice (see

Appendix A3). These findings suggest that even though children's ability to control and plan their motor movements does predict their ability to produce recognizable drawings, this factor alone does not fully account for the observed developmental changes.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.690	0.173	-3.979	<0.001
Age (in years)	0.251	0.020	12.805	<0.001
Est. drawing frequency	-0.062	0.173	-0.356	0.721
Average tracing score	0.267	0.020	13.529	<0.001
Time spent drawing	0.039	0.021	1.868	0.062
'Ink' used	-0.031	0.020	-1.546	0.122
Number of strokes	0.008	0.018	0.477	0.634
Age*drawing frequency	0.017	0.017	1.042	0.298

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. All predictors were z-scored so that coefficients are comparable.

Recognizable drawings contain more category-diagnostic information across development. The above results suggest that children gradually improve their ability to include diagnostic visual information in their drawings. However, they are also consistent with an alternative account where younger children are just as able to produce recognizable drawings when they are engaged with the task, but are less likely to stay on task than older children and thus produce unrecognizable drawings more often. 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 visual information 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 drawings that were correctly classified to contain similar amounts of diagnostic visual information, regardless of whether they were produced by younger or older children.

We found that even for drawings that were correctly classified (38.6% of the balanced subset of drawings, N=8,590), the amount of diagnostic information they contained increased as a function of age, as measured by the log-odds probability assigned by the logistic-regression classifier to the target category (see *Methods*, $B = 0.111$, $SE = 0.015$, $df = 3544.18$, $t = 7.354$, $P < 0.001$) (see Appendix, Table A6. This analysis provides converging evidence that age-related improvements in children's abilities to produce recognizable drawings reflect a gradual increase in the amount of category-diagnostic information in their drawings.

Unrecognizable drawings still contain semantically relevant information. Even if a child does not know the diagnostic features of *giraffes* or *rabbits*, they likely know that both are animals with four legs. Thus, this kind of coarse semantic information may still be contained in children's unrecognizable drawings. Indeed, prior work suggests that basic-level recognition – recognizing something as a *rabbit* – is not a pre-requisite for inferring semantic information. For example, adults can reliably judge the animacy (animate vs. inanimate) and real-world size of unrecognizable, textured images by inferring that animals tend to have high curvature and that larger, inanimate objects (e.g., couches) tend to have boxier shape structures (Long, Konkle, Cohen, & Alvarez, 2016; Long, Störmer, & Alvarez, 2017; Long, Yu, & Konkle, 2018) and children appear to be sensitive to these cues by the preschool years (Long, Moher, Carey, & Konkle, 2019).

Following this idea, we reasoned that even young children's children's misclassified drawings might contain information about the animacy and real-world size of the category they were intending to draw (see Figure 3A). To examine this possibility, we analyzed the patterns of misclassifications made by the logistic regressions and found that misclassified drawings reliably carried information about both real-world size (see

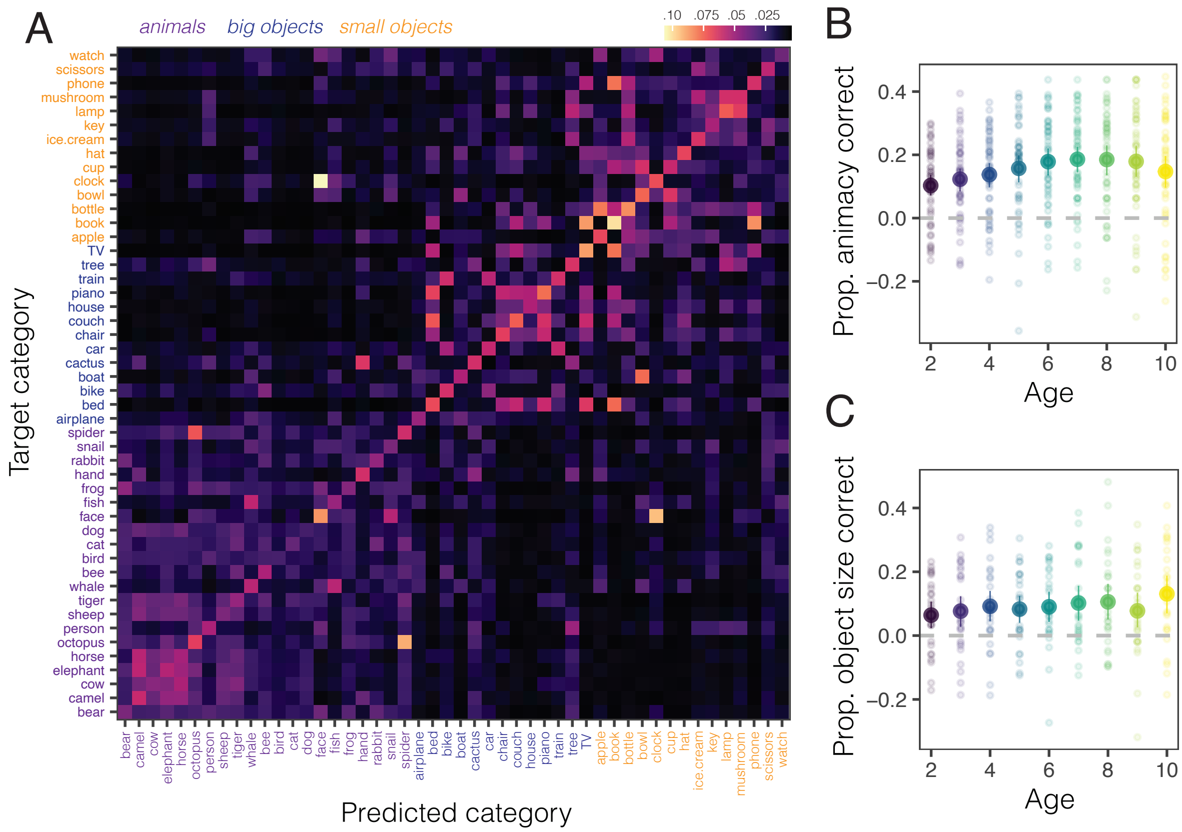

Figure 3. A: Classifier probabilities for the subset of drawings that were misclassified on the basis of VGG-19 embeddings (FC6). The y-axis shows the category children were intending to draw; the x-axis shows all of the categories in the dataset. Lighter values represent greater classifier probabilities assigned to a given category (see colorbar). (B,C). Proportion of misclassified drawings that contained the correct animacy/object size information of the target category (relative to baseline in the dataset). 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.

Figure 3C) and animacy (see Figure 3B) across all ages. We also found substantial structure in the pattern of probabilities assigned by the classifier to the other categories (see Figure 3A): for example, unrecognizable drawings of *an octopus* were often assigned a high classifier probability for a *a spider*. These results suggest that children's unrecognizable drawings are far from meaningless scribbles; instead, they contain relevant semantic information about the category children were intending to draw.

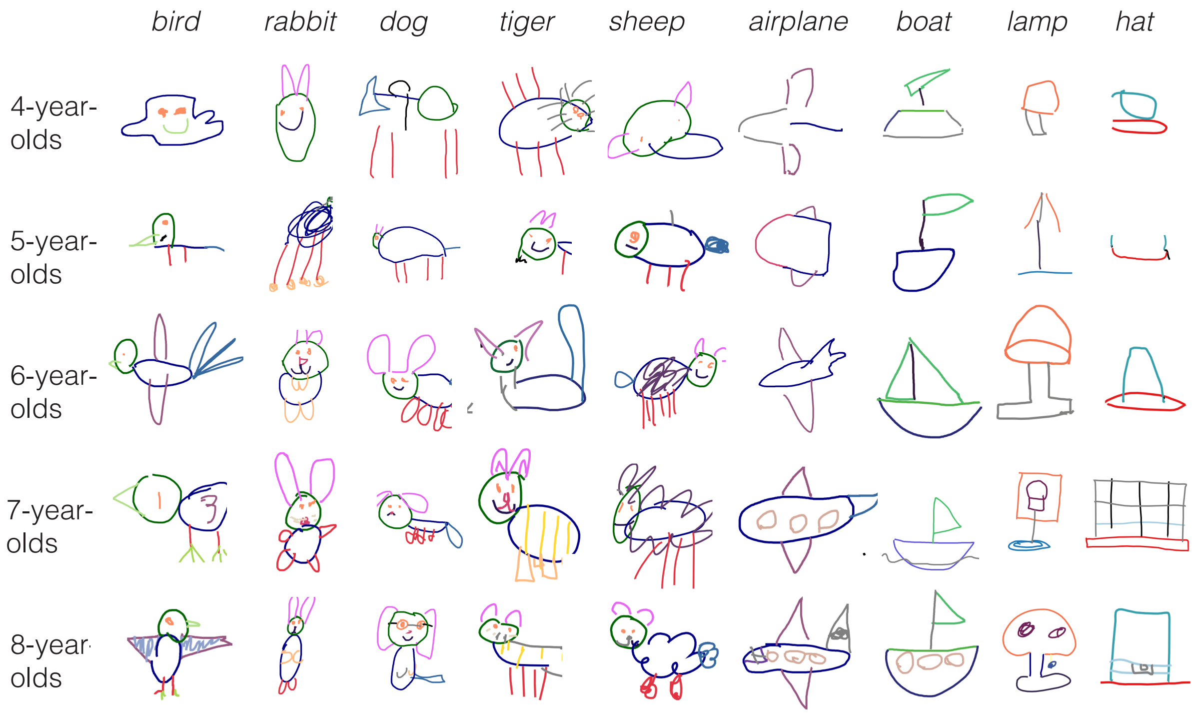

Figure 4. Example drawings from 4-8 year old's, with part annotations. Each color represents object parts labels agreed upon by human annotators; grey lines represent strokes with multiple parts, and black lines represent unintelligible strokes.

Drawings contain different semantic parts across development. While these findings so far provide strong evidence for age-related gains in the ability to produce recognizable drawings, it is not clear exactly what aspects of these drawings account for this improvement. A natural possibility is that children gradually learn which object parts to include and how much to emphasize those parts (e.g., long *ears* for rabbits) in their drawings (Tversky, 1989; Tversky & Hemenway, 1984) as they learn about the semantic properties of those categories. To explore this idea, we collected semantic part annotations of the visible object parts in a subset of N=2021 drawings, and examined developmental changes in which parts children prioritized in their drawings throughout development.

Consistent with this idea, we found that drawings produced by older children generally contained more unique semantic parts than drawings by younger children (B = 0.395, SE = 0.041, df = 2071, P < .001; Figure 5), lending support to the notion

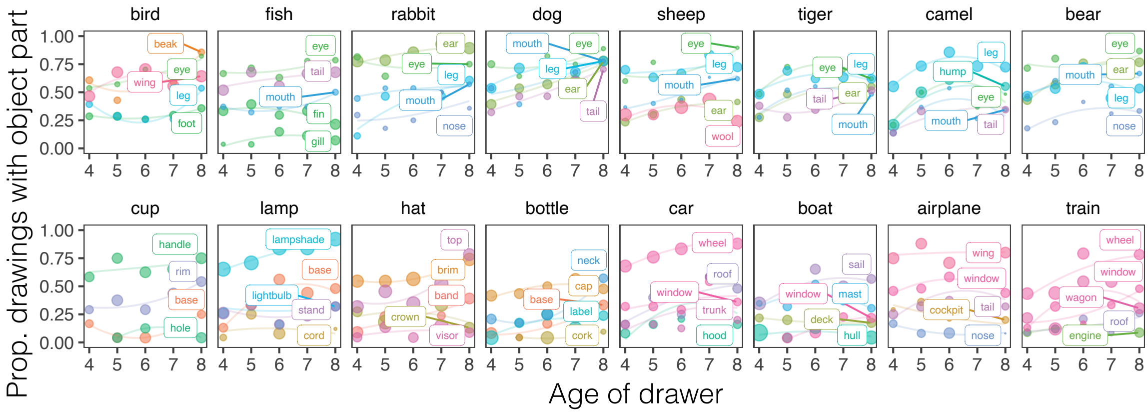

Figure 5. Proportion of drawings that included a given object part for each object category as a function of children's age (see *Methods*). The size of each dot reflects the average emphasis (proportion of stroke length relative to the entire drawing) for each object part within each bin (max plotted part emphasis = .5); the top five most frequent object parts are included for each category excluding generic "body/head" parts.

that, across development, children learn to more effectively express what they know about various visual concepts by enriching their drawings with additional semantic part information (see Appendix, Figure A8). For some concepts, these gains appeared to be specific to single part: for example, older children were more likely to produce *cups* with *handles* and *cars* with recognizable *wheels*. For other concepts, however, age-related changes were more complex: for example, in drawings of *bears*, both *ears* and *eyes* appeared to change in prevalence and emphasis. Further, while most younger children's drawings of *rabbits* included recognizable *ears*, which are in principle informative about the category, many of them were still not recognizable as *rabbits*. Taken together, these exploratory findings suggest that while there are clear age-related changes in the part complexity of children's drawings, the mere presence of – or amount of emphasis on – any particular part may not be sufficient to account for developmental variation in its recognizability (see Appendix Figure A9). Instead, they suggest that visuospatial information about what these parts look like and how they are arranged may be needed to explain why drawings by older children are more recognizable than those by younger ones. For example, *ears* on *rabbits* may need to be more elongated relative to the *head* to provide a strong enough cue to category membership.

Development of drawing recognition

Why do children include more diagnostic visual information in their drawings as they grow older? One source of these developmental changes may be refinements in children's internal visual concepts. As children acquire more knowledge about the visual distinctions between visual concepts, children might more clearly represent the visual information that best distinguishes depictions of *rabbits* from *dogs*, for example, and may be able to use this information when recognizing drawings. This account thus predicts that children should improve over development in their ability to exploit visual information in drawings to recognize their intended meaning.

To explore 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 thus varied in the degree to which they were recognizable and hence 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 information.

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 visual concepts (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*).

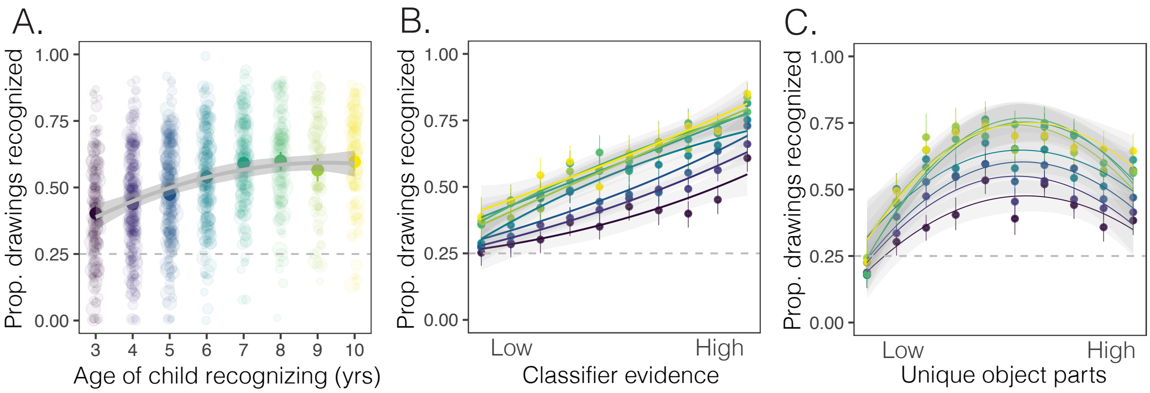

Figure 6. A. Drawing recognition as a function of the age of the child who participated in the guessing game; each dot represents data from one child who participated and is scaled by the number of trials they completed. (B,C). Drawing recognition data plotted separately by the age of the child participating as a function of the (B) amount of diagnostic visual information in each drawing, operationalized as the the *classifier evidence* assigned to each sketch relative to the distractor categories and (C) the number of unique object parts in each drawing. Both variables are binned into deciles for visualization purposes. Error bars represent bootstrapped 95% confidence intervals.

Overall, we found that children became steadily better at identifying the category that a drawing referred to (see Figure 6A). In contrast, performance on photograph matching trials was relatively similar across ages. All children whose data were included in our analyses scored >75% correct on photograph 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.

Children's drawing recognition improves over development. We next evaluated the idea that children's ability to exploit category-diagnostic visual information during recognition improves over childhood. We first tested how children's drawing recognition ability varied with respect to the amount of diagnostic visual information in a given drawing. To do so, for each drawing that appeared in the guessing games, we measured diagnostic visual information. We fit a 4-way logistic regression classifier trained on the VGG-19 features extracted from the drawings presented in each guessing game (see *Methods)* and measured diagnostic information as

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.050	0.121	0.412	0.680
Classifier evidence	0.477	0.046	10.405	<0.001
Recognizer Age	0.317	0.019	16.777	<0.001
Classifier evidence*Recognizer Age	0.062	0.014	4.246	<0.001

Table 2

Model coefficients of a GLMM predicting visual recognition performance as a function of recognizer age and classifier evidence. All predictors were z-scored such that coefficients are comparable.

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 models predicting children's recognition performance with child's age, this metric of diagnostic visual features, and their interaction as fixed effects (see *Methods* for further details and robustness checks using CLIP).

Drawings with more diagnostic visual information were better recognized across all ages (see Table 2, Figure 6B, CLIP robustness check in Appendix, Table B2). Yet older children were also better able to capitalize on graded differences in the diagnostic visual information in drawings when recognizing them (see Figure 6B), evidenced by an interaction between classifier evidence and recognizer age in both cases. This result held when we restricted our analyses to a subset of high-performing children who performed at ceiling on photograph matching trials (see *Appendix*, B1) suggesting that these effects are unlikely to be driven by a differences across development in either engagement or in the ability to match drawings with the picture-cue buttons (see individual category effects in *Appendix*, B3). Children became steadily better over development at using diagnostic visual information to recognize the intended meaning of line drawings.

Children's drawing recognition varies with part information. Finally, we evaluated how children's abilities to use object part information during visual recognition changes across development (Juttner et al., 2016; Mash, 2006). Specifically, we examined how children's recognition accuracy varied with the number of unique recognizable object parts included in the drawings they tried to recognize. We again fit a generalized linear mixed effect model to the recognition data, modeling the interaction between the number of unique parts in each drawing and the age of the child recognizing the drawing, with the maximal random effects structure supported by our data.

We found that drawings with more unique object parts tended to be better recognized – but, unexpectedly, that drawings with many object parts were less well recognized than drawings with an intermediate number of parts (see Figure 6C) though there was substantial variation across categories (see Appendix, Figure A1, see Table B3). Critically, we again found that older children were better able to capitalize on increasing object part information during recognition, as evidenced by a significant interaction between between the number of unique parts in the drawings and the age of child recognizing the drawings. In other words, children's ability to integrate additional object part information during recognition changed across development (see all coefficients in Appendix, Table B3). These additional analyses using object part annotations suggest that children's ability to use diagnostic visual information during recognition matures gradually throughout childhood.

Relationship between visual production and recognition

So far, our descriptive results suggest that both visual production and recognition of drawings develop gradually and in parallel throughout childhood. To what degree do these developmental trajectories reflect changes in the same mental representations of visual concepts? Insofar as children's abilities to recognize what drawings mean and to produce meaningful drawings both rely on a shared mental representation, then their ability to produce a drawing of a *dog* may be correlated with their ability to recognize a drawing of a *dog*, as in adults (Perdreau & Cavanagh, 2014). To explore this possibility, we borrow an analysis technique used in language acquisition (Braginsky, Yurovsky, Marchman, & Frank, 2019), where variation in word production is well-predicted by

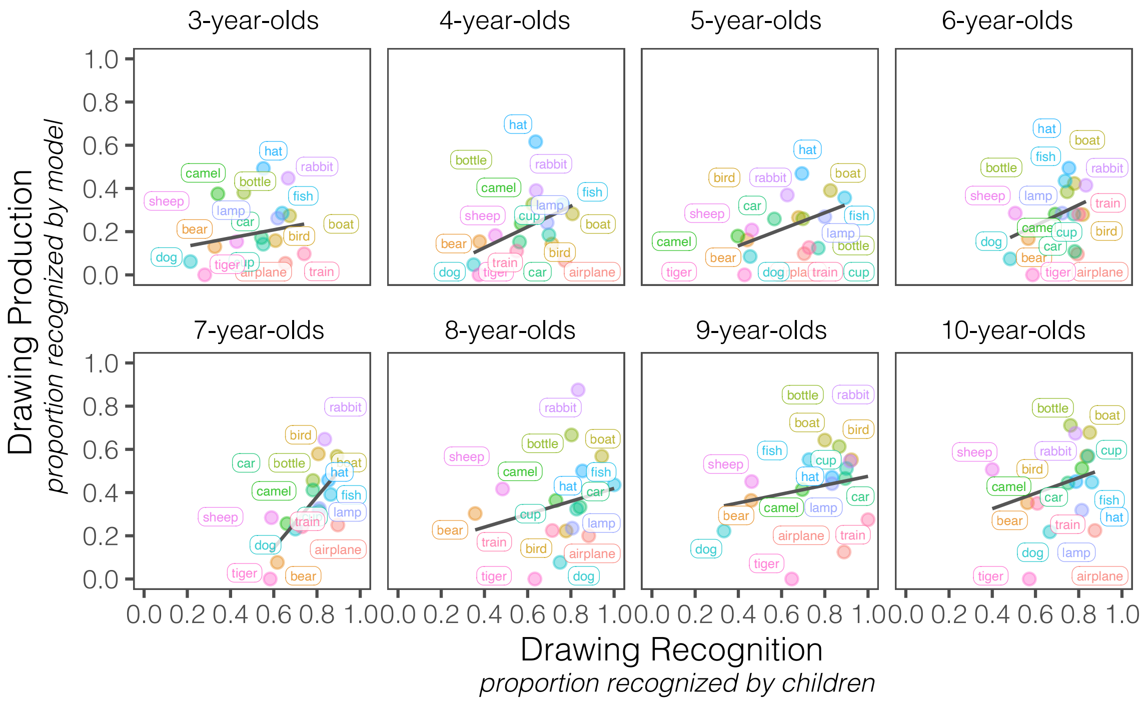

Figure 7. Each dot represents a category (e.g., hat) at a given age (in years), where the y-axis value represents how well children of that age produced recognizable drawings of that category (as assessed by CLIP model classifications) and the x-axis value represents how well children of that age were able to *recognize* the top 30% most recognizable drawings of that category (as assessed by accuracy in the 4AFC recognition games). Independent sets of drawings are analyzed in each case.

independent data about that word, e.g., the frequency of a word in English books (Goodman, Dale, & Li, 2008).

We thus explored how well variation in visual production is related to visual recognition at the category level, acknowledging the exploratory nature of these analyses.[2] To do so, we examined children's visual production and recognition abilities using independent sets of drawings of the same categories. To estimate drawing recognition ability, we used children's performance on the guessing games to calculate

[2] While some children may have both contributed drawings and participated in the recognition game with different categories, these sessions were anonymous, and thus we do not have access to within-child data.

how often children of a given age, on average, were able to recognize drawings of a given category. To ensure that we were examining children's visual recognition abilities for relatively recognizable drawings, we analyzed how well children could recognize the top 30 percent most recognizable drawings of each category (using CLIP classification probabilities, see *Methods*). To estimate drawing production abilities, we calculated how often children at each age could produce recognizable drawings of a given category (e.g., *a dog*), as assessed by CLIP model recognition scores. For these exploratory analyses, we used CLIP model classifications, as CLIP showed less dramatic category variation relative to VGG-19 classifications (see Appendix, Figure A2). In addition, within the independent set of sketches used to assess children's recognition, CLIP showed a higher correlation with children's recognition behaviors (aggregating across individual sketches; VGG-19, r=.28; CLIP, r=.43).

Overall, we found that children's visual production and visual recognition abilities were positively related at the category-level at all ages (see Figure 7). For example, while *dogs* and *sheep* were consistently both harder to produce and to recognize, *rabbits* and *hats* were easier to produce and recognize. Thus, these exploratory results suggest relative consistency across categories in these two tasks, suggesting that children's ability to perform well in both tasks may rely on a shared visual representation, and paving the way for future work that seeks to understand the sources of this category variation using within-child, controlled experiments.

General Discussion

We conducted a systematic investigation of how children produce and recognize line drawings of a wide range of everyday visual concepts across development (2-10 years of age). We developed a large dataset of children's digital drawings (>37K) and capitalized on innovations in machine learning to quantify changes in children's drawings across development. We found robust improvements in children's ability to include diagnostic visual information via recognizable object parts in their drawings, and these developmental changes were not reducible to either increased effort or better visuomotor abilities. Further, we found that children's unrecognizable drawings still contained information about the animacy and real-world size of the visual concepts they were trying to depict; highlighting an intermediate stage between scribbles and fully recognizable drawings. We also found improvements throughout childhood in children's ability to recognize each other's drawings, particularly in their ability to capitalize on diagnostic visual information during drawing recognition tasks. Together, these results document parallel developmental changes in how children use diagnostic visual information when producing and recognizing freehand line drawings, suggesting that refinements in children's visual concepts may underlie improvements across both tasks. As children's perceptual abilities (Bova et al., 2007; Natu et al., 2016), semantic knowledge (Tversky, 1985; Vales, Stevens, & Fisher, 2020), and visuomotor skills (Li & James, 2016) evolve across childhood, children's changing visual concepts influence both how children produce and recognize freehand line drawings.

More broadly, the present work highlights how the combination of modern machine-learning methods and larger-scale datasets of naturalistic behaviors can contribute to theoretical progress 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 across categories. In turn, our use of innovations in computer vision and computational modeling allow the analysis of the entirety of this large dataset, capturing variation across both unrecognizable and recognizable drawings in a single analytic approach (which would have been intractable with human ratings, e.g., how similar each *dog* drawing was to every other category in the dataset). Using this approach, we were able to distinguish variability in children's drawings due to a range of different developmental processes – including motor skill and task effort – from variability related to visual concept knowledge. 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 (Karmiloff-Smith, 1990; Minsky & Papert, 1972) both using large-scale datasets and controlled, within-child experiments that directly relate visual production and recognition and examine item variation.

Variation across visual concepts

Our exploratory analyses suggested that children's abilities to produce and recognize drawings were correlated at the category-level at each age, e.g, drawings of *dogs* were both harder to produce and to recognize. Further, estimates of how often these items tend to be drawn or experienced did not explain this variation. Why might some categories be easier to draw and to recognize?

One possibility is that these item effects are related to other metrics of visual experience with different categories beyond what we have measured: Perhaps exposure frequency in educational materials or children's media were not adequately captured by our surveys. Or perhaps children may experience more invariant exemplars of some categories, making it easier to identify and draw those categories. For example, while *dogs* may vary substantially, *ice cream cones* have a relatively more invariant form. Children may in turn develop more refined visual representations for more frequently experienced and more invariant items, leading to more recognizable drawings.

Any effects of exposure frequency or form variability might interact with the degree to which a category has a 3D shape structure that can be easily depicted using a two-dimensional line drawing. For example, canonical *mushrooms* have a relatively simple shape, whereas *rabbits* have many sub-parts that need to be depicted (i.e., legs, ears, nose, mouth, tail), and children may also struggle to arrange these sub-parts in a way that conveys the meaning of the visual concept (i.e., the correct ratio between the size of the rabbit's ears and head). In turn, these shape structures may lead different categories to have more or less typical iconic representations that children are accessing when producing and recognizing line drawings. For example, *trains* are often depicted as steam trains, as modern trains can be hard to distinguish from other vehicles as line drawings. Future work may be able to use our dataset to examine the relationship of these factors to visual production and recognition, as we purposefully included items that vary along these dimensions.

From scribbles to categories

The present work also highlights the gradual progression in children's drawings from exploratory scribbles through an intermediate stage (Morra & Panesi, 2017) where their drawings may not unequivocally convey a specific visual concept (e.g., a *giraffe*), while still containing enough visual information to be recognizable as an "animal." An intriguing implication is that the mechanisms by which children reliably produce such semantically ambiguous drawings might be related to mid-level visual representations in the brain that are sensitive to coarse distinctions between broad classes of visual objects (i.e., large inanimate vs. small inanimate vs. animals) without relying on category-level distinctions (Long et al., 2016, 2019, 2017). Moreover, drawing tasks may allow children to convey this kind of partial knowledge about a visual concept that may otherwise be difficult for them to express verbally. And as children learn more about a specific visual concept – for example, that *giraffes* have longer legs than *antelopes* or *elephants* – these incremental gains in conceptual knowledge may manifest in their drawings, even if they are not yet clearly recognizable as a*giraffe*. This work thus showcases the potential of drawing production tasks for examining graded changes in how children's knowledge about visual concepts grows and changes across development.

Possible learning mechanisms

Several learning mechanisms are consistent with the developmental changes we observed. One possibility is that children become better visual communicators as they learn which visual 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. Indeed, both drawing experts (Perdreau & Cavanagh, 2013b) and naïve adults who practice drawing similar categories (Fan et al., 2020) show better enhanced visual recognition abilities of these categories. 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). 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, *ice cream cones* were among the best recognized categories and estimated (by parents) to be among the least practiced by children. In addition, we did not see any obvious changes in the developmental trajectory around 6 years of age when most children start to write (though this achievement might cause smaller changes that we could not detect in our data); rather, we observed evidence for gradual growth throughout the entire age range.

A second, non-exclusive possibility is that children are explicitly learning the diagnostic features of categories as they enrich their semantic knowledge. For example, children may learn about the functional properties of different attributes: *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 percolate into children's visual concepts and thus be 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, again not mutually exclusive with the other two, is that children are implicitly learning category-diagnostic information through the process of visual categorization itself: through repetitively viewing and categorizing depictions, real-life examples, and photographs of these different categories. Indeed, the neural networks used here to categorize drawings did not have visuomotor experience drawing or training about the semantic properties of these categories. Thus in principle it is possible that children could be refining their visual concepts without substantial involvement from other cognitive or sensorimotor systems.

Limitations and future directions

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 (Henrich, Heine, & Norenzayan, 2010). Children in different contexts may spend considerably more or less time viewing and producing depictions of different categories, and thus could vary reasonably affect how how they represent them. Further, different cultural contexts have different conventions for depicting visual concepts, as evidenced by drawings from adults (Lewis, Balamurugan, Zheng, & Lupyan, 2021). However, there are likely to be some aspects of drawing production and interpretation that are broadly shared across cultural contexts (Cavanagh, 2005; Hertzmann, 2020), given prior work that has investigated picture comprehension in communities with modest exposure to Western visual media (Deregowski, 1989; Kennedy & Ross, 1975). Moreover, there is evidence from earlier work that some of this convergence may reflect evolutionarily conserved visual processing mechanisms, as non-human primates exhibit can recognize the correspondence between line drawings and their real-world referents (Itakura, 1994;

Tanaka, 2007). Future work that examines drawings across different cultural contexts in both adults (Lewis et al., 2021) and children will help quantify the consistency and variability in how we represent and depict visual concepts.

Second, while we imposed strong filtering requirements, we were not present while the children were drawing or guessing at the kiosk 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. Nonetheless, we replicated our main experimental results on drawing production in a controlled, experimental context with a smaller set of categories (see Appendix, Figure A4).

Third, since these datasets are cross-sectional, they do not directly relate visual production and recognition abilities at the individual level. Our exploratory, category-level analyses suggest variation in these two abilities are correlated across development; ultimately, however, within-child measurements will be necessary to confirm that changes in children's visual concepts underlie the observed changes in both tasks. In addition, these correlational analyses can only provide hints as to whether changes in visual production cause changes in visual recognition or vice versa. Finer-grain, within-child training studies (as in Bremner & Moore, 1984) could provide traction on the direction of causality between visual production and recognition.

Conclusion

Our results call for further 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, and we hope that the open datasets and tools we have created here will open up new avenues for such future work. We propose that a full understanding of how children produce and recognize drawings of visual concepts 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; no other demographic information was collected. 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 (see 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*); these categories were also chosen to overlap with those in the QuickDraw database of adult drawings (https://github.com/googlecreativelab/quickdraw-dataset). Each set of category prompts that rotated at the station thus included both animate and inanimate categories as well as commonly and infrequently drawn categories; category prompts were presented in a random order.

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. When observing participants interacting with the drawing station, we noted that most children's parents did not fill out the survey because they were either talking to other parents or taking care of a sibling. Further, while children could contribute drawing data more than once if they chose, this did not occur during our structured observation of the kiosk. This protocol was approved by both the Institutional Review Board at Stanford University (43992, Development of Children's Drawing Abilities).

Raw drawing data 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 = 13119 drawings after exclusions (15.8% 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.

Experimental Dataset Procedure

In a separate experiment, children were seated in front of a touchscreen tablet with a trained experimenter. As in the larger dataset, children completed two shape-tracing trials, and then children produced drawings of 12 familiar object categories (a*irplane, bike, bird, car, cat, chair, cup, hat, house, rabbit, tree, watch*) which were randomly assigned to different cue-types (verbal vs. picture). In this paper, we analyze only verbal-cued drawings for sake of comparison to the drawing station dataset. 135 children participated in the experiment; 6 participants were excluded, (3) for skipping more than 6 drawing trials and (3) for scribbling three or more times in a row. Six additional participants were tested but their data was not recorded due to a technical error, and two participants never advanced past the practice trials, leading to a final sample of 121 children. No additional demographic data was recorded about the participants. This protocol was approved by the Institutional Review Board at Stanford University (43992, Development of Children's Drawing Abilities).

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

VGG-19: 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. For our main analysis, 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. For additional analyses using the earlier convolutional layers, we first applied spatial averaging over the outputs of each layer to reduce their dimensionality, as in Fan et al., 2018, before also applying channel-wise normalization.

VGG-19: Logistic regression classifiers. Next, we used these features to train object category decoders. 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 visual information that is 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.

CLIP Classifications. For these analyses, we used the ViT-B/32 implementation of CLIP publicly available at https://github.com/openai/CLIP. Model features were extracted for center-cropped versions of each sketch in the entire dataset (N=37770), and for the tokenized, text versions of the labels for each of the 48 categories (e.g. "a dog"). We then computed the cosine similarity between the features for each sketch and each of the 48 category labels and assessed which category label received the highest similarity. If the category label that had the highest similar was the category children were prompted to draw, this was counted as a correct classification.

Human recognition scores: Experimental Dataset. We measured the recognizability of each drawing in the controlled, experimental dataset via an online recognition experiment. Adult participants based in the U.S. were recruited via Prolific for a 15-minute experiment and asked to identify the category depicted in a random subset of approximately 140 drawings; each drawing was shown to 10 participants. Participants were shown these drawings in a random sequence and asked "What does this look like?" and selected their responses from the set of 12 categories and were encouraged to provide their best guess if they were unsure. No participants were excluded from analysis for missing the catch trial, which was included to verify that participants could accurately describe their goal in this task. We then computed a recognition score for each drawing, reflecting the proportion of participants who correctly identified the target category.

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 misclassified drawing, we calculated whether the category assigned by the logistic regression classifier was 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) (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 4-8 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-, and 8-year-olds). If children completed the entire session, this resulted in a total of 48 trials for each participant (40 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's 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 registrations on https://osf.io/qymjr/).

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.

Semantic part annotation task

Crowdsourcing category part decompositions. We designed a web-based crowdsourcing platform and recruited 50 English-speaking adult participants from Prolific to identify the basic parts of objects for each of the 16 object categories. On each trial, participants were cued with a text label of an object category and asked to list 3 to 10 object parts that came to mind (e.g., head, leg, tail, etc. for "tiger"). Participants were instructed to write only concrete parts of an object (e.g., "tail") rather than abstract attributes (e.g., "tufted"), to use common names of parts rather than technical jargon (e.g., "prehensile"), and to generate as complete a part list as they could for each object category. We applied lemmatization to the resulting part decompositions to remove redundant part labels, such as "hoof" and "hoofs", and manually edited part labels that were spelled incorrectly or with alternative spellings. We then selected the top 10% of part names that were most frequently listed. This generated a total of 82 object parts with a range of 5-13 possible parts per object category.

Part labeling task. We developed a web-based annotation paradigm adapted from previous research (Huey, Walker, & Fan, 2021; Mukherjee, Hawkins, & Fan, 2019) to obtain detailed annotations of how each pen stroke in children's drawings corresponded to the different parts of the depicted objects. 1,034 English-speaking adult participants were recruited from Prolific and completed the semantic annotation task. We excluded data from 78 additional participants for experiencing technical difficulties with the web interface (N=11) and for having low accuracy on our attention-check trial (N=67). Data collection was stopped when every drawing had received annotations from at least three annotators.

Each annotator was presented with a set of 8 drawings randomly sampled from the drawing dataset but consistent within the same animacy and object size (i.e., small animals, large animals, vehicles, household objects). Each drawing was accompanied by the name of its object category (e.g., "airplane"), as well as a gallery of crowd-sourced part labels that corresponded to it. For each stroke in the presented drawing, annotators were prompted to tag it with the part label that described the part of the depicted object that it represented. Annotators were permitted to label a stroke with multiple part labels if they believed a stroke to represent multiple different parts of the depicted object, and were able to write their own custom label if they believed that none of the provided part labels were fitting. They could also label a stroke as unintelligible if they could not discern what it represented. Annotators also completed an "attention-check" trial, consisting of a pre-selected drawing that had been annotated by a researcher and then randomly inserted into the set of drawings. If annotators did not match the researcher's annotation criteria for this drawing, data sessions from these annotators were excluded from subsequent analysis.

Annotation data preprocessing. First, we evaluated how often annotators agreed on what each stroke of children's drawings represented by calculating the inter-rater consistency among annotators. Across drawings, annotators agreed on the same part label for 69.9% of strokes. There was modest improvement in agreement across age, with with drawings produced by older children eliciting more consistent annotations (4-year-old drawings = 68.3% mean agreement, 8-year-old drawings = 69.8% mean agreement). We retained stroke annotations that were assigned the same part label(s) by at least two of three annotators. While annotators infrequently wrote custom labels (we did not analyze custom annotations for the present analysis) they only used 68 of the available 82 part labels. Our resultant dataset therefore contained 14,159 annotated strokes across 2,088 drawings.

Part inclusion and emphasis calculation. For part inclusion, we calculated the number of unique object parts assigned to each drawing; strokes labeled as unintelligible were not counted as distinct parts. For part emphasis, we calculated the proportion of the total length of strokes that were attributed to a particular object part in a drawing (e.g., wings), relative the total length of all strokes in the entire drawing (including strokes that were not agreed upon or that were unintelligable). If strokes were used to represent multiple object parts, we took the total length of the stroke and divided it by the number of parts that it was assigned to.

Data Availability

Source data are provided with this paper. The drawings and pre-processed data that support the findings are available at https://osf.io/qymjr/

Code Availability

The code used analyze the data are available at https://osf.io/qymjr/

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 on this project throughout several years and Megan Merrick for assistance with data collection. This work was funded by an NSF SPRF-FR Grant #1714726 to BLL, a NIH K99 HD108386 to BLL, and a Jacobs Foundation Fellowship to MCF.

References

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature communications*, *10*(1), 1–13.

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.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, *11*(1), 1–14.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

BLINDED. (in press). Developmental changes in drawing production under different memory demands in a u.s. and chinese sample. *Developmental Psychology*. Retrieved from BLINDED

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.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., … Hodges, J. R. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive neuropsychology*, *20*(1), 27–47.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

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*.

Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, *434*(7031), 301–307.

Chamberlain, R., Drake, J. E., Kozbelt, A., Hickman, R., Siev, J., & Wagemans, J. (2019). Artists as experts in visual cognition: An update. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(1), 58.

Chamberlain, R., Kozbelt, A., Drake, J. E., & Wagemans, J. (2021). Learning to see by learning to draw: A longitudinal analysis of the relationship between representational drawing training and visuospatial skill. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 76.

Clottes, J. (2008). *Cave art.* Phaidon London.

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.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2022–03.

Cox, M. V., & Ralph, M. L. (1996). Young children's ability to adapt their drawings of the human figure. *Educational Psychology*, *16*(3), 245–255.

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., Pierroutsakos, S. L., & Uttal, D. H. (2003). The origins of pictorial competence. *Current Directions in Psychological Science*, *12*(4), 114–118.

Deregowski, J. B. (1989). Real space and represented space: Cross-cultural perspectives. *Behavioral and Brain Sciences*, *12*(1), 51–74.

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.

Finke, R. A., & Slayton, K. (1988). Explorations of creative visual synthesis in mental imagery. *Memory & cognition*, *16*(3), 252–257.

Fisher, A. V., Godwin, K. E., & Matlen, B. J. (2015). Development of inductive generalization with familiar categories. *Psychonomic bulletin & review*, *22*, 1149–1173.

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*.

Freeman, N. H., & Janikoun, R. (1972). Intellectual realism in children's drawings of a familiar object with distinctive features. *Child development*, 1116–1121.

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

Gibson, J. J. (1971). The information available in pictures. *Leonardo*, 27–35.

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.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515-531.

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29.

Hertzmann, A. (2020). Why do line drawings work? a realism hypothesis. *Perception*, *49*(4), 439–451.

Huey, H., Walker, C. M., & Fan, J. E. (2021). How do the semantic properties of visual explanations guide causal inference? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Itakura, S. (1994). Recognition of line-drawing representations by a chimpanzee (pan troglodytes). *The Journal of general psychology*, *121*(3), 189–197.

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.

Kennedy, J. M., & Ross, A. S. (1975). Outline picture perception by the songe of papua. *Perception*, *4*(4), 391–406.

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.

Kozbelt, A. (2001). Artists as experts in visual cognition. *Visual cognition*, *8*(6), 705–723.

Lewis, M., Balamurugan, A., Zheng, B., & Lupyan, G. (2021). Characterizing variability in shared meaning through millions of sketches. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

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

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., Moher, M., Carey, S. E., & Konkle, T. (2019). Animacy and object size are reflected in perceptual similarity computations by the preschool years. *Visual Cognition*, *27*(5-8), 435–451.

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.

Luquet, G.-H. (1927). Le dessin enfantin.(bibliothèque de psychologie de l" enfant et de pédagogie.).

Mash, C. (2006). Multidimensional shape similarity in the development of visual object classification. *Journal of Experimental Child Psychology*, *95*(2), 128–152.

Minsky, M., & Papert, S. (1972). *Artificial intelligence progress report* (Tech. Rep.). Cambridge, MA, USA.

Morra, S., & Panesi, S. (2017). From scribbling to drawing: the role of working memory. *Cognitive Development*, *43*, 142–158.

Mukherjee, K., Hawkins, R. X., & Fan, J. W. (2019). Communicating semantic part information in drawings. In *Cogsci* (pp. 2413–2419).

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*.

Perdreau, F., & Cavanagh, P. (2013a). The artist's advantage: Better integration of object information across eye movements. *i-Perception*, *4*(6), 380–395.

Perdreau, F., & Cavanagh, P. (2013b). Is artists' perception more veridical? *Frontiers in neuroscience*, *7*, 6.

Perdreau, F., & Cavanagh, P. (2014). Drawing skill is related to the efficiency of encoding object structure. *i-Perception*, *5*(2), 101–119.

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*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

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.

Rosch, E. (1978). Principles of categorization.

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

Sayim, B. (2011, October). What line drawings reveal about the visual brain. , 1–4.

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.

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.

Tanaka, M. (2007). Recognition of pictorial representations by chimpanzees (pan troglodytes). *Animal cognition*, *10*(2), 169–179.

Tchalenko, J. (2009). Segmentation and accuracy in copying and drawing: Experts and beginners. *Vision research*, *49*(8), 791–800.

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

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, *25*(6), 983.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of experimental psychology: General*, *113*(2), 169.

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.

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

Supplemental analyses: visual production across childhood

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 drawing dataset by each age group.

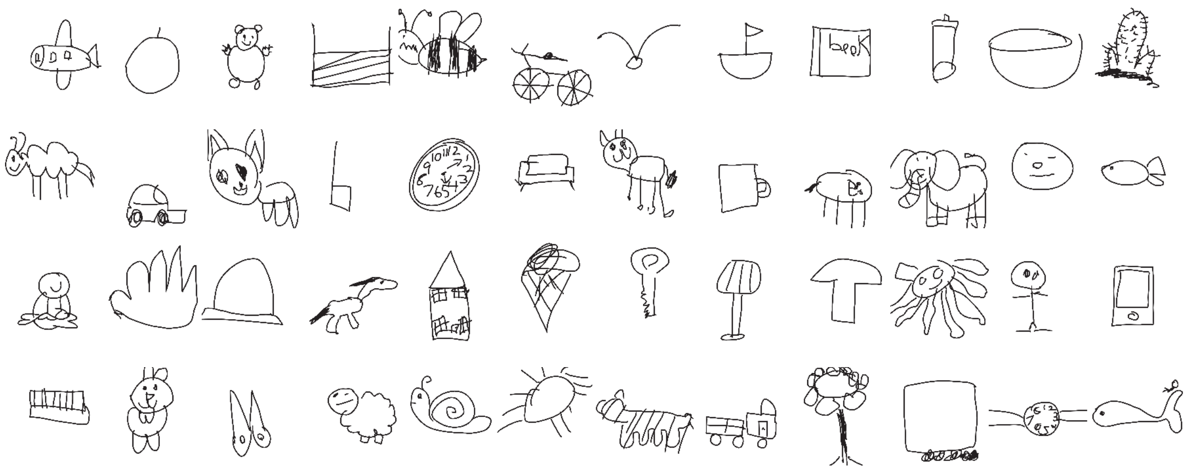

Figure A1. Examples of correctly classified drawings from each of the 48 categories presented at the experiment station in alphabetical order: airplane, apple, bear, bed, bee, bike, bird, boat, book, bottle, bowl, cactus, (2nd row): camel, car, cat, chair, clock, couch, cow, cup, dog, elephant, face, 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.

	Term	VIF	SE_factor
1	Age	1.26	1.12
2	Drawing Frequency	1.00	1.00
3	Tracing score	1.24	1.11
4	Draw duration	1.30	1.14
5	Ink used	1.29	1.14
6	Number of strokes	1.07	1.03
7	Age* Drawing frequency	1.01	1.00

Table A2

Results of the multicolinearity analysis for the predictors used in the main GLMM predicting the recognizability of children's drawings.

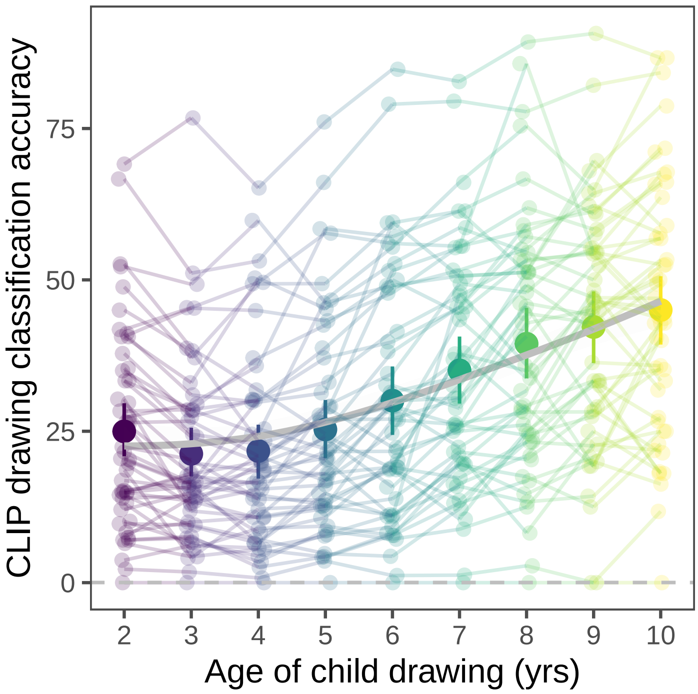

Figure A2. Figure 2A, redone with CLIP classifications. Y-axis shows classification accuracy as a function of children's age (x-axis). Each dot represents data from an individual category, which are connected by individual trend lines. Error bars represent bootstrapped 95% confidence intervals.

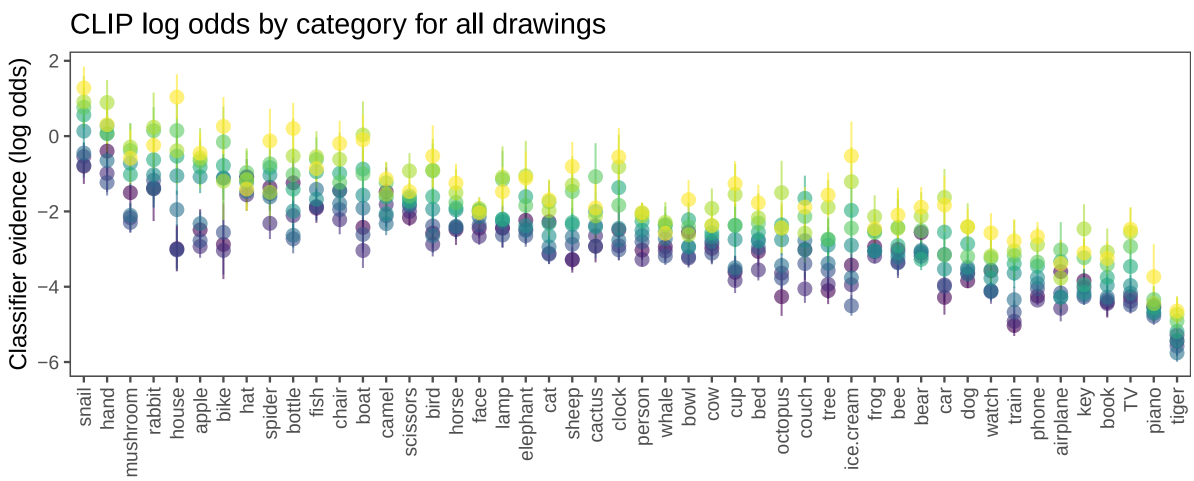

Figure A3. CLIP log-odds probabilities (y-axis) assigned to each category as a function of children's age; each dot represents data from an individual category and age.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.319	0.178	-7.410	<0.001
Age (in years)	0.329	0.019	17.225	<0.001
Est. drawing frequency	0.274	0.177	1.551	0.121
Tracing score	0.279	0.020	14.320	<0.001
Time spent drawing	0.195	0.019	10.065	<0.001
'Ink' used	0.047	0.018	2.642	0.007
Number of strokes	0.070	0.030	2.338	0.019
Age*drawing frequency	0.029	0.014	2.030	0.042

Table A3

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

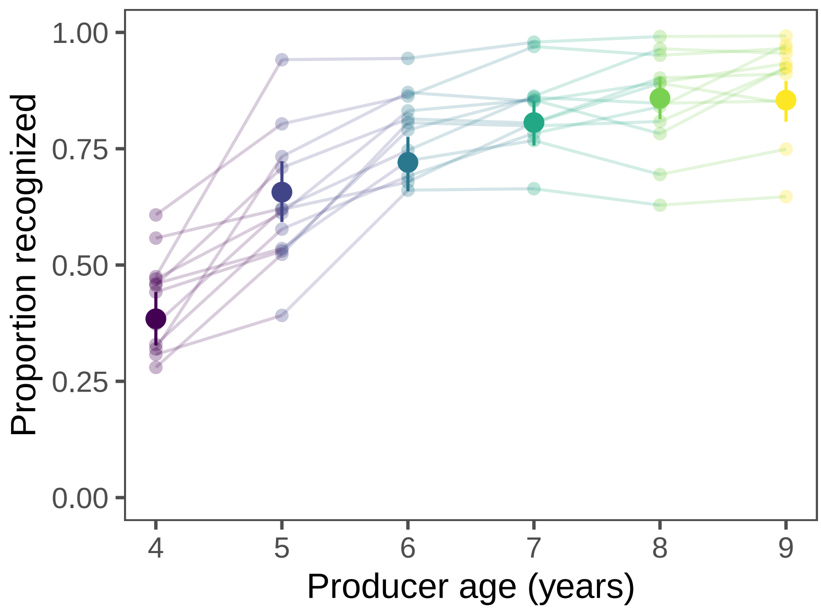

Figure A4. Drawing accuracy as a function of children's age; children drew in response to verbal prompts in a controlled, experimental setting. Y-axis reflects the proportion of human observers who correctly identified the drawing in a 12AFC guessing task. Error bars reflect 95 percent bootstrapped confidence intervals.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.736	0.024	12.391	30.226	<0.001
Drawing frequency	0.043	0.024	9.853	1.833	0.097
Age (in years)	0.151	0.011	247.928	14.348	<0.001
Drawing frequency * Age	0.007	0.007	1282.435	1.073	0.283

Table A4

Model coefficients of a linear mixed effect model predicting the recognizability of each drawing (as assessed by crowd-sourced adult behavioral data). Drawings were produced in an experimental context. All predictors are z-scored and random intercepts for each category and participant are included.

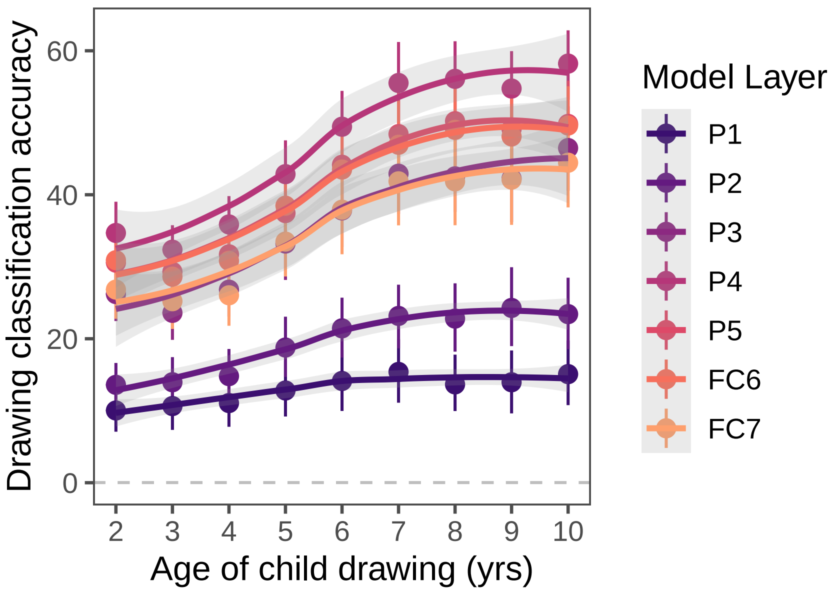

Figure A5. Drawing accuracy as a function of children's age using embeddings from each layer in the VGG-19 network. Error bars represent 95 percent bootstrapped confidence intervals.

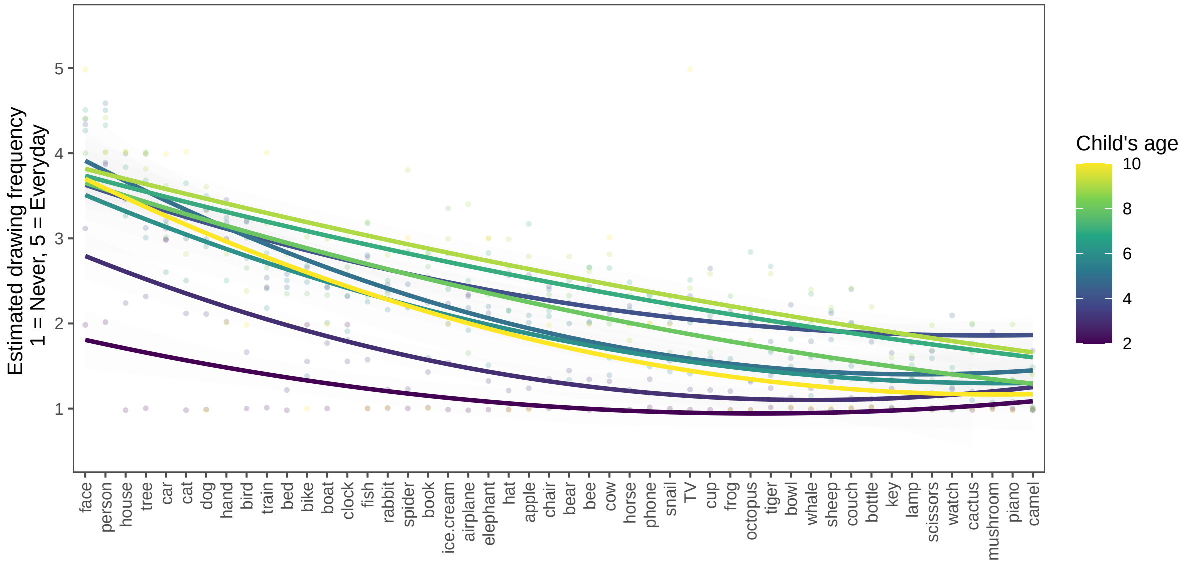

Figure A6. Frequency (y-axis) with which parents (recruited online) estimated their children drew each of the 48 categories in the dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.965	0.237	-4.065	<0.001
Children's age	0.360	0.021	17.006	<0.001
Freq. in adult books	0.515	0.312	1.653	0.098
Est. AoA	0.702	0.418	1.678	0.093
Freq. in CHILDES	0.129	0.398	0.323	0.747
Drawing frequency	-0.290	0.326	-0.889	0.374

Table A5

Model coefficients of a GLMM predicting the recognizability of each drawing (i.e. binary classification scores) from children's age, the frequency of each category in CHILDES, the estimated Age-of-Acquisition, and the frequency of each word in adult English books. All predictors are z-scored; random intercepts for each category and participant are included.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.687	0.107	43.356	-6.434	<0.001
Children's age	0.111	0.015	3544.182	7.354	<0.001
Drawing frequency	0.020	0.114	42.904	0.174	0.863
Average tracing rating	0.101	0.015	3964.435	6.753	<0.001
Time spent drawing	-0.019	0.018	7710.462	-1.060	0.289
Ink spent	0.018	0.017	7888.865	1.042	0.297
Number of strokes	0.063	0.018	8110.693	3.535	0.000
Age*Drawing frequency	0.011	0.013	7672.028	0.811	0.417

Table A6

Model coefficients of a linear mixed effect model predicting the log-odds probability assigned to correctly classified drawings using VGG embeddings, including random intercepts for each category and participant.

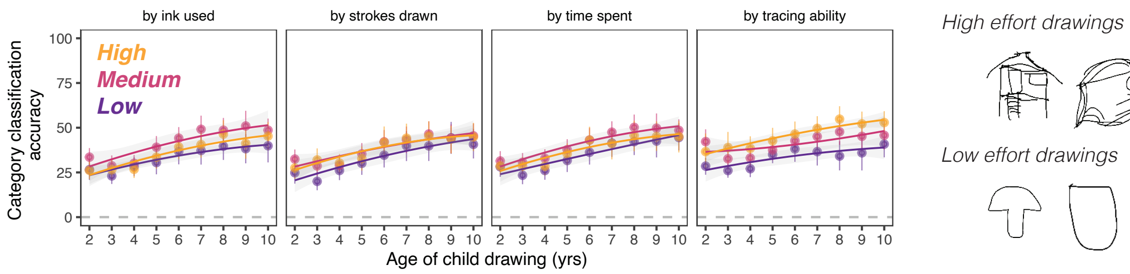

Figure A7. (Left): Classification accuracy by age, split into bins according to whether children expended a greater/lesser amount of strokes, ink, or time, and by their estimated tracing abilities (see Methods). (Right): Example drawings where children spent higher/lower amounts of *effort*—greater/lower than average number of strokes, time spent drawing, or 'ink' used.

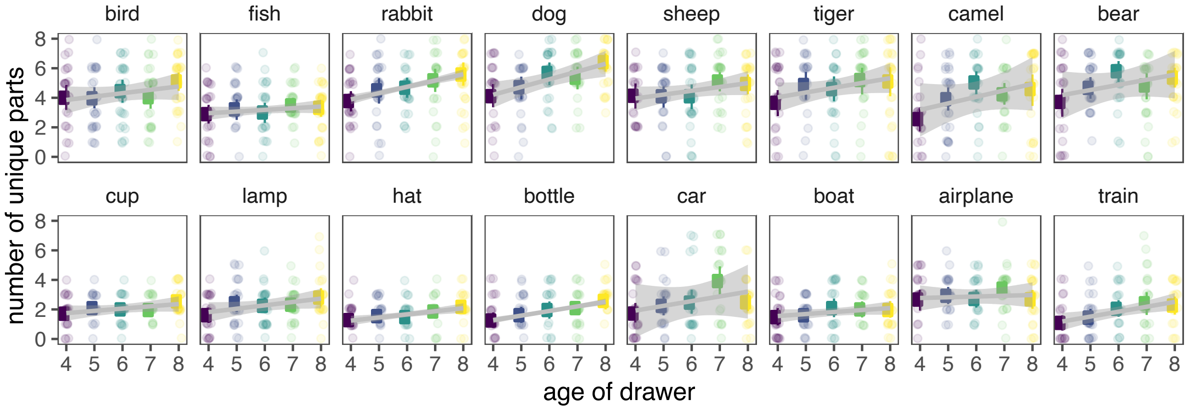

Figure A8. Number of unique parts per each object category included in the semantic part annotation subset (N=2,088 drawings of 16 categories) across age.

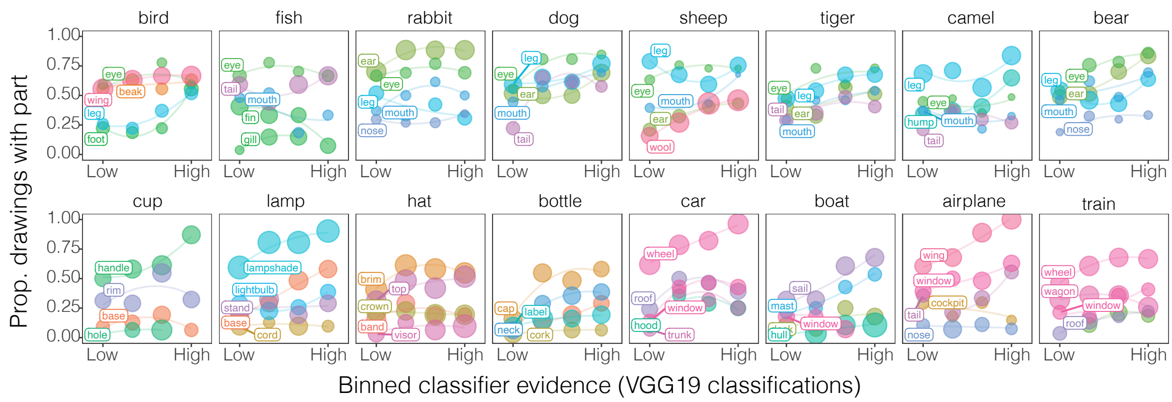

Figure A9. Object parts included for each category as a function of the VGG-19 classification evidence, binned into quartiles; as in the main texts, only the top 4 object parts that were frequently included (excluding head/body) are shown here. Dot size represents the visual emphasis on each part.

Appendix B

Supplemental materials: visual recognition

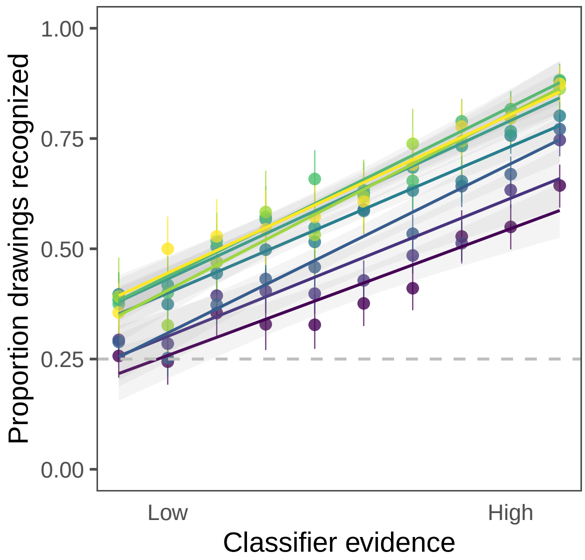

Figure B1. Replication of the main interaction on visual recognition behaviors (proportion recognized, y-axis) by recognizer age (individual lines colored by age) and classifier evidence, here using CLIP classification probabilities (binned into deciles on the x-axis).

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	Pr(>|z|)
(Intercept)	0.668	0.099	6.715	<0.001
Classifier evidence (scaled)	0.518	0.051	10.057	<0.001
Recognizer age (scaled)	0.141	0.023	6.190	<0.001
Classifier evidence*Recognizer Age	0.056	0.023	2.464	0.014

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.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.397	0.183	2.167	0.030
Classifier evidence (scaled)	0.987	0.215	4.590	<0.001
Recognizer age (scaled)	0.367	0.021	17.644	<0.001
Classifier evidence*Recognizer Age	0.091	0.018	5.079	<0.001

Table B2

Model coefficients of a GLMM predicting visual recognition performance as a function of recognizer age and the category-diagnostic information in drawings (derived from CLIP embeddings, see Methods).

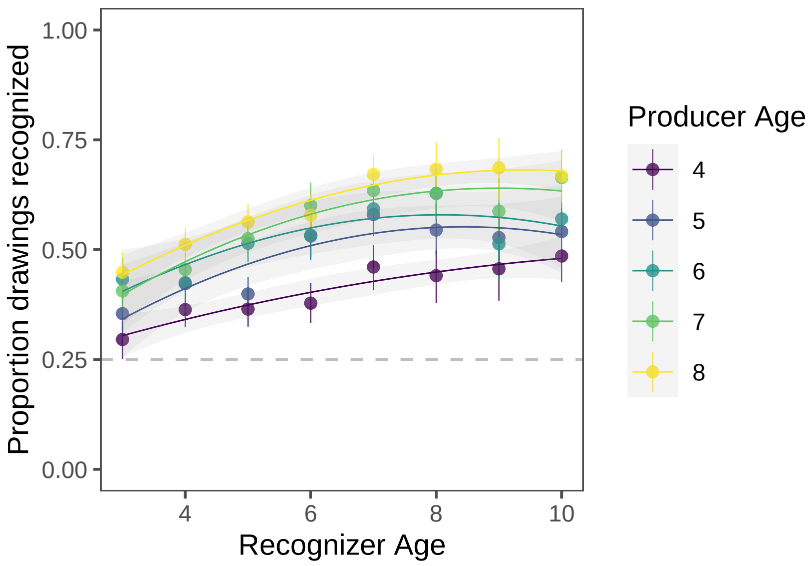

Figure B2. Proportion of drawings recognized (y-axis) as a function of both the age of the child participating (x-axis) and the age of the child who originally produced the drawing (each line represents a different age). Error bars depict 95% bootstrapped confidence intervals.

	Estimate	SE	z value	Pr(>|z|)
(Intercept)	0.05	0.23	0.20	0.84
Parts	129.39	10.04	12.89	<0.001
Parts**2	-34.98	3.24	-10.79	<0.001
Age	0.34	0.02	17.04	<0.001
Parts x Age	12.70	2.71	4.69	<0.001
Parts**2 x Age	-8.95	2.46	-3.64	<0.001

Table B3

All model coefficients from a generalized, linear mixed effect model predicting how well children could recognize drawings of visual concepts as a function of their own age (Age; recognizer age) and the number of unique parts included in each drawing.

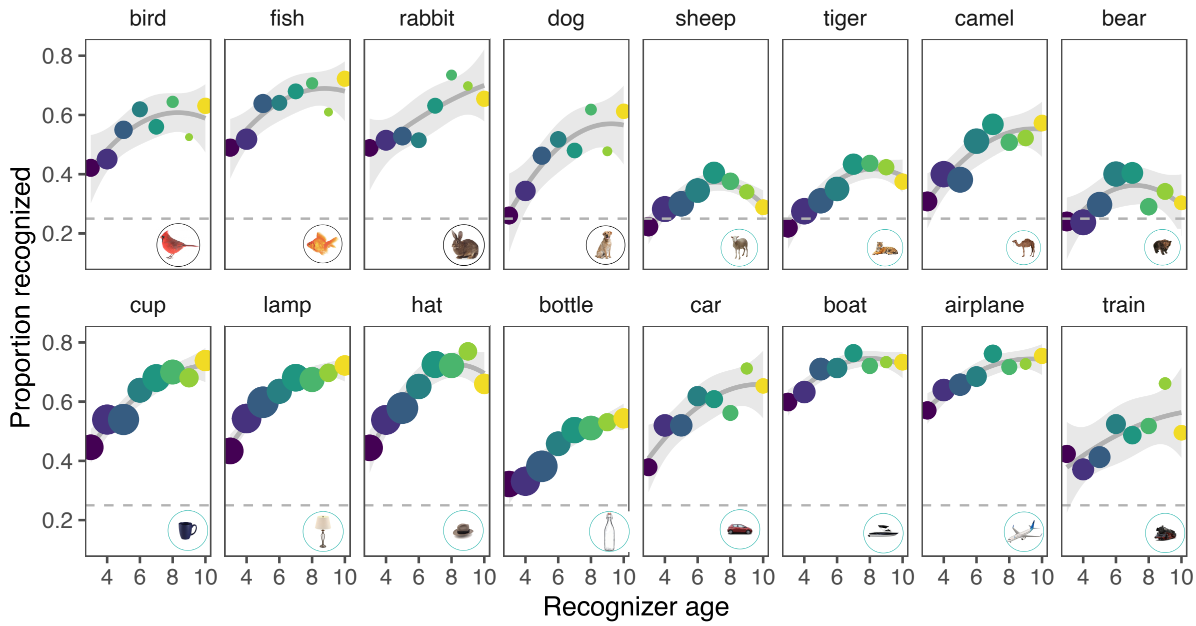

Figure B3. Children's drawing recognition behavior for each of the 16 categories included in the recognition games; categories are grouped by the respective 4AFC game they were embedded in. Dots are scaled by the amount of data available from each age for each category (younger children were more frequent participants). Photo icons for each category are shown in the bottom right of each panel.

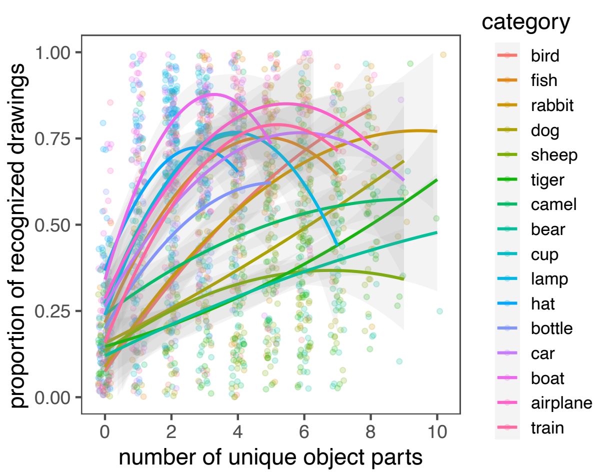

Figure B4. Drawing recognition for each category as a function of the number of unique parts included in each drawing; each individual dot is a unique drawing.