Developmental changes in the precision of visual concept knowledge

Bria Long[1] & Ernst-August Doelle[1,2]

[1] University of California San Diego

[2] Stanford University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Bria Long: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Bria Long, 9500 Gillman Drive, La Jolla, CA 92093. E-mail: brlong@ucsd.com

Abstract

How precise is children's visual concept knowledge, and how does this change across development? We created a gamified picture-matching task where children heard a word (e.g., "swordfish") and had to choose the picture "that goes with the word." We collected data from large sample of children on this task, and we modeled changes in the proportion of children who chose a given image for a certain word over development. We found gradual developmental changes in children's ability to identify the correct category. Error analysis revealed that children were more likely to choose higher-similarity distractors as they grew older; children's error patterns were increasingly correlated with CLIP target-distractor similarity. These analyses suggest a transition from coarse to finer-grained visual representations over early and middle childhood. Broadly, these findings demonstrate the utility of combining gamified experiments and similarity estimates from computational models to probe the content of children's evolving visual representations.

Keywords: visual concepts, receptive vocabulary, large language models, object recognition

Word count: X

Developmental changes in the precision of visual concept knowledge

```
## Warning: package 'kableExtra' was built under R version 4.3.2
```

Introduction

When a child hears a word — like a “whale” — this activates a mental representation of its referent in the real-world. But what is this representation actually like? Depending on how old a child is—and how much they have learned about whales—they might imagine a canonical exemplar of a blue whale, a specific whale from a picture book, or perhaps they just know vaguely that a whale is an animal that lives in the ocean. How precise are the visual representations that underlie children’s understandings of words across childhood?

As infants begin to communicate with their caregivers, they experience an astonishing rate of vocabulary growth (Bloom, 2000; Braginsky et al., 2022). Infants as young as 6-months of age appear to absorb some shape information from label-object co-occurrences in everyday experience (Vong et. al 2024; Bergelson et al., 2009), and 14-18 month-olds extend newly learned words to atypical exemplars (Weaver et al., 2024, Child Dev). By around their second birthday, children extend words to stylized, 3D exemplars (Smith, 2003) as they learn that shape is a valuable cue to basic-level categories (Rosch et al., 1976). Thus, at least for within-category exemplars, very young children exhibit relatively sophisticated generalization abilities for common visual concepts, in line with a broad-to-narrow view of category development (Waxman & Gelman, 2009), where infants construe words as initially referring to many items and subsequently refine their representations across development. From this perspective, children’s visual representations may change relatively little across childhood; instead, children may gradually acquire new visual concepts and instead change in how they represent the relationships between visual concepts: for example, children may learn that whales are mammals, and then appropriately group them with other land mammals vs. with fish when asked to make taxonomic classifications. Accordingly, empirical work on children's developing ability to recognize objects (Azyenberg & Behrman, 2024) has also focused on the first few years of childhood as the most critical period in which object recognition abilities develops.

Contrary to this simplified account, here we posit that children's visual concepts change throughout childhood, with an extended developmental trajectory that continues in parallel with later vocabulary learning. Of course, children's vocabulary knowledge—often assessed via paper-and-pencil, closed, expensive traditional assessments—grows and expands across childhood (CITE), but there has been relatively little consideration of the visual representations that support children's performance on picture vocabulary tasks. As children enter schooling environments and begin to learn why animals and objects are classified the way they are, this semantic learning is likely to influence the visual features that are prioritized in children's visual concepts. Some work on children's production and recognition of drawings of common objects hints at this kind of protracted developmental timeline (Long et al., 2024): in a large observational study, children became increasingly able to both depict and recognize line drawings of common object categories. However, no work has directly tested children's visual recognition behaviors for a wide variety of visual concepts. We suspect that this is in part because of the difficulty of obtaining data from large samples of children on a consistent set of items with variability over a large developmental age range.

To overcome these methodological barriers, we created a gamified picture-matching task where children heard a word (e.g., "swordfish") and had to choose the picture "that goes with the word". Critically, we chose distractor items with high, medium, and low concept similarity to each target word; distractors were paired via cosine similarity of the target and distractor words in a large multimodal language model (CLIP, Radford et al., 2021). This task was then deployed in online, preschool, and school contexts to 3599 children aged 3-15 years and 211 adults years of age. Using this large dataset, we find gradual changes in how children represent visual concepts across childhood, with older children becoming both more accurate at identifying the correct referents throughout this extended age range; however, we also found that even young children were more likely to choose the related vs. unrelated distractors, highlighting a gradual change from coarse, representations that encompass both the target and related distractors to fine-grained, specific representations that the visual information that words refer to. We then use both unimodal and multimodal embeddings from this same modelto examine how visual, linguistic, and multimodal similarity explain children's error patterns across development.

Methods

Procedure

Children were invited to participate in a picture matching game; a cover story accompanied the game where children were asked to help teach aliens on another planet about some of the words on our planet; children were able to pick a particular alien to "accompany them on their journey." Before the stimuli appeared, children heard a target word (e.g., "apple") and then were asked to "choose the picture that goes with the word". The four images appeared in randomized locations on the screen, and one of the images always corresponded to the target word. On practice trials, the distractor images were all very dissimilar to the target concept, and the target word was realtively easy. The tablet played a chime sound if they chose correctly, and a slightly unpleasant sound if they responded incorrectly. Each child viewed a random subset of the item bank, and the items they viewed were displayed in a random order. Children were allowed to stop the game if they wanted to. While different versions of the game included varying amounts of trials or items, as these games are part of a larger project to develop an open-sourced measure of children's vocabulary knowledge as an alterantive to the PPVT; however, here we analyze children's responses to items that were generated using the THINGS+ dataset with distractors of varying difficulty.

Participants

To obtain a large sample, we collected data from children in several different testing contexts. We collected data from children in an in-person preschools (N = 65, 3-5 year-olds), from the Children Helping Science Platform, (N=243, 3-7 year-olds), 6 elementary schools, and 9 charter schools across multiple states (N=3332, 5-14 year-olds) and adults online (N=211 adults, recruited via Prolific; half of the adults spoke English as a second language). Most participants responded via a touch-screen tablet, except those recruited online: however, children's parents responded via clicking on the image on Children Helping Science, and adults responded via clicking on the images.

After pre-processing, we included for a total of from 3786 participants from preschools, schools, and online testing contexts around the country (range 84 to 654), who completed, on average, 25.02 4AFC trials that were sampled randomly from the stimuli set (max=86; different maximum numbers of trials were included in different testing contexts). We tested an additional 84 participants who scored near chance on 4AFC trials (chance=25%, threshold=30%) and were school-aged (>6 years of age) and who we excluded from analyses; these participants completed an average of 17.72 trials.

Stimuli selection

We capitalized on publicly available existing image and audio databases to generate stimuli. Visual concepts were taken from the THINGS+ dataset (Stoinski et all., 2023), after filtering out non-child safe images (e.g., weapons, cigarettes) and images wtih low nameability, as per the released norming data. We used the copy-right free, high-quality image released for each visual concept. We then subset to visual concepts that had available audio recordings in the MALD database as well as age-of-acquisition (AoA) ratings from a previous existing dataset (Kuperman, 2012).

Using this subset, we sampled distractors with high, medium, and low similarity to the target word as operationalized via embedding similarity of the words in the language encode of a multimodal large language model (CLIP, Contrastive Language-Image Pre-training, Radford et al., 2021). High-, medium, and low similarity values were determined relative to the distribution of possible target-distractor pairing values for each word in the THINGS+ dataset. In our final set, we had 108 items with a range of different estimated age-of-acquisitions (e.g., hedgehog, mandolin, mulch, swordfish, waterwheel, bobsled) with all unique targets and distractors; in addition, we constrained the sampling such that target-distractor pairs had estimated age of acquisition within 3 years of each other. All stimuli and their meta-data are available on the public repository for this project.

Model features

We obtained all model features features using the OpenAI available implementation of CLIP available at https://github.com/openai/CLIP. For language similarity, we computed the cosine similarity of the embeddings of the target to each distractor word on each trial (e.g,. rose – tulip, rose – glove, rose – hubcap). For visual similarity, we repeated this procedure but by obtaining image similarity vectors in the vision transformer. For multimodal similarity, we computed the cosine similarity of the embedding of the target word in the language model to the embeddings for each of the distractor images; this is possible because the embedding spaces for the vision and language transformers in the CLIP model are aligned and have the same number of dimensions.

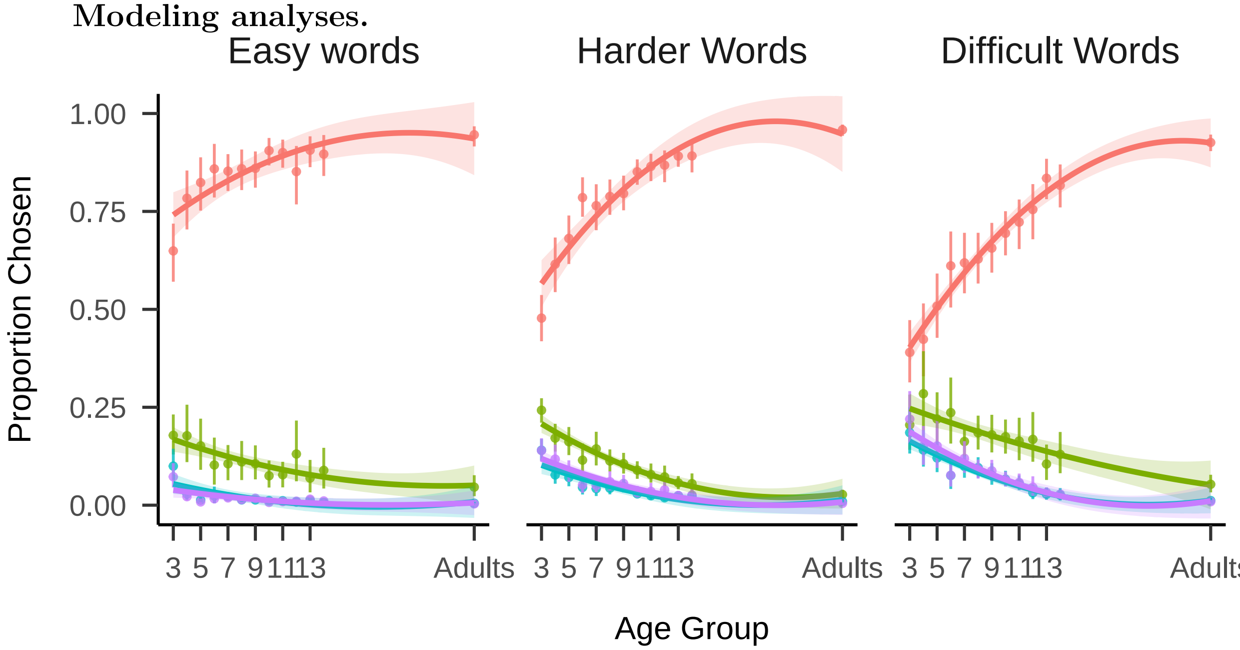
Results

Growth in visual concept knowledge across age

Partial knowledge in young children

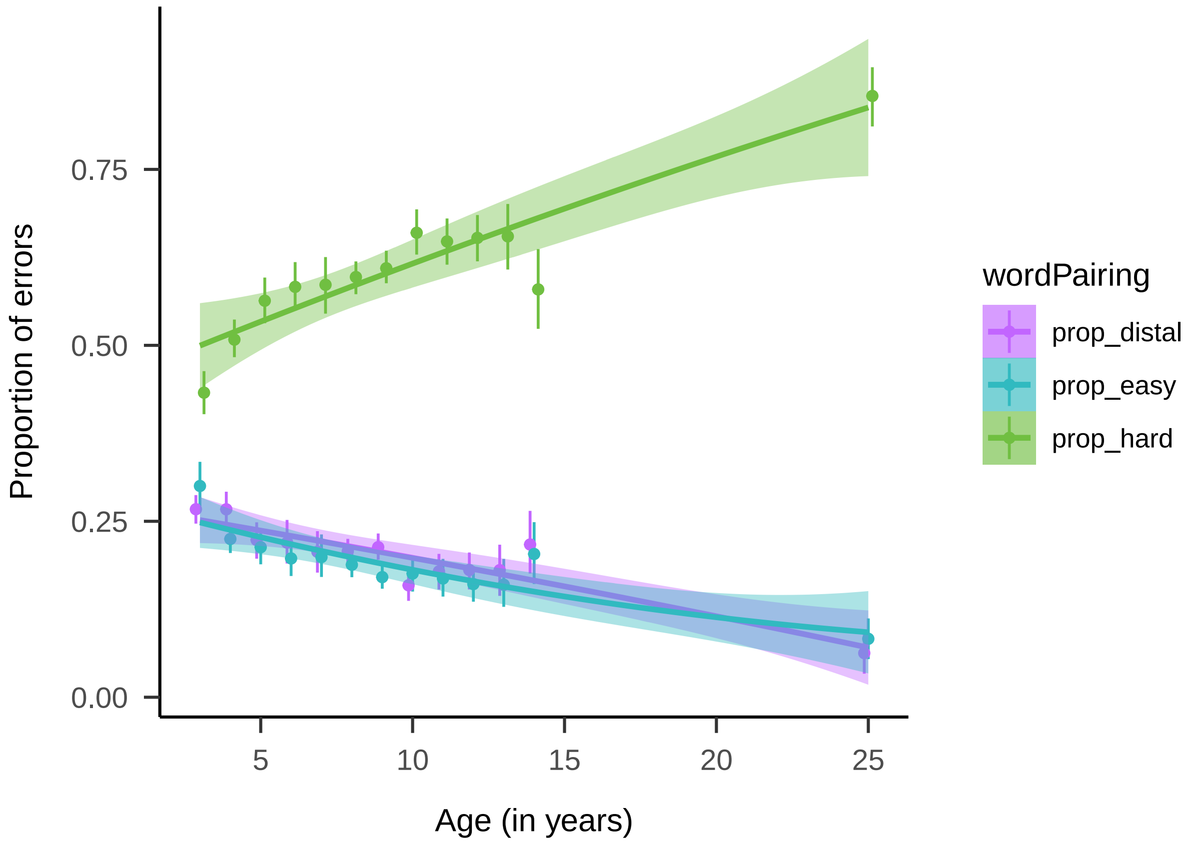
Changes in precision of visual concepts

Modeling analyses.



Older children are more likely to choose related distractors

construct CIs by age and make individual data structures for plotting




```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(hard, total_num_errors) ~ scale(age_group) + (1 | pid)
##    Data: error_by4afc_for_glmer
## Control: glmerControl(optCtrl = list(maxfun = 20000), optimizer = c("bobyqa"))
##
##      AIC      BIC   logLik deviance df.resid
##   8610.2   8628.7  -4302.1   8604.2     3423
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4464 -0.3782  0.1340  0.4211  1.3540
```

```
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  pid    (Intercept) 0        0
## Number of obs: 3426, groups:  pid, 3409
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.56493    0.01336 -42.287  < 2e-16 ***
## scale(age_group)  0.08603    0.01177   7.312 2.64e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## scal(g_grp) 0.380
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

Discussion

How precise is children's visual concept knowledge, and how does this change across development?

Overall, these analyses suggest a transition from coarse to finer-grained visual representations over early and middle childhood.

Children's visual concept knowledge gradually becomes more refined as children learn what distinguishes similar visual concepts from one another. Broadly, these findings demonstrate the utility of combining gamified experiments and similarity estimates from computational models to probe the content of children's evolving visual representations.

Implications: Supports Ecological enrichment accounts:

On another viewpoint, there is also substantial enrichment and change in children's visual representations of everyday visual concepts. Broader view on the learning environment (e.g., Bruner), and children as quite active participants in their learning environment, Longer view on the timeline for learning, which in turn changes how we think about the relevant learning environment— which changes substantially as they grow and learn both from their peers throughout early childhood and in structured educational contexts. For example, they may have grossly misrepresented the sizes of certain objects (e.gg., whales are XX bigger than dolphins) and certain visual features may become more or less salient as they understand their functional roles (XX) or semantic relevance of the category. On this account, even school-aged children's visual representations may undergo substantial change as they learn more about the world around them, even as their vocabulary growth tapers. Goes beyond acquisition account to suggest that their representations change beyond what has been measure in classic recognition tasks with young children]

Connection to adult expertise We suspect that visual concept learning extends into adulthood, and that many adults have coarse visual representations for many different words. Consider that while we experience the referents of some visual concepts relatively frequently—e.g., trees, computers, cups, cars—other words refer to visual concepts that different individuals may have varying amounts of interest in and frequency in interacting with—like telescopes, or antelopes. Visual concept learning is likely influenced by both children and adults' occupation and pre-occupations. And indeed decades of work has established that birding experts, car aficionados, and graphic artists have both qualitatively and quantitatively different kinds of visual representations for the visual concepts that they engage with (CITE, CITE):

References

Table 1

Fixed Effects from Generalized Linear Mixed Effects M

effect	Predictor	*b*	*SE*	*z*	*p*
fixed	Intercept	-0.56	0.01	-42.29	< .001
fixed	Age (scaled)	0.09	0.01	7.31	< .001

Note.

Analysis conducted using a generalized linear mixed effects model with a binomial distribution. The mo