Developmental changes in the precision of visual concept knowledge

Abstract

How precise are the visual representations that underlie children’s understanding of words, and how does this precision change across development? We assessed children’s visual vocabulary knowledge in a game where children heard a word (e.g., “swordfish”) and had to choose a matching picture. Distractors were chosen to vary in their linguistic similarity to each target word in a multimodal language model. We collected data from a large sample of children (*N* = 3575, 3–14 year olds) and adults (*N* = 211 adults) and found gradual changes over developmental age ranges in participants’ ability to identify the image that a word referred to. When children made errors, younger children were more likely to choose unrelated distractors than older children; error patterns were best explained by combining linguistic and visual similarity. These results highlight that children have partial knowledge about many visual concepts and document a transition from coarse to finer-grained representations.

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

*Word count:* 1999 excluding Method/Results

**Research Transparency Statement**

**General Disclosures:** Conflicts of interest: All authors declare no conflicts of interest. Funding: This research was supported by [blinded]. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article, beyond the generation of text and image embeddings for stimuli selection and analyses. Ethics: This research received approval from a local ethics board at [blinded], IRB numbers (IRB-67514, IRB-58960, IRB-20009, and IRB-19960).

**Main Study.** These analyses presented in the paper were not pre-registered, and children’s data were analyzed both during and at the end of data collection. These data were collected with the additional goal of creating an open-source measure of children’s developing vocabulary knowledge. The experiment code and stimuli are available at <https://osf.io/ynkg4/?view_only=6d823174f6514ff589ba3567ffe1992b>. All pre-processed data and analysis and plotting code have been made available at the same public online repository. The raw, trial-by-trial data cannot be shared due to data sharing agreements between the participating schools and the institution’s IRB as part of their research partnership agreements.

Developmental changes in the precision of visual concepts

Across the world’s languages, many words refer to visible things in the environment (e.g., “*thermos*”, “*prism*”, “*whale*”). How precise are the visual representations that underlie children’s understanding of these words—children’s *visual concepts*? For example, depending on how old a child is (and how much they know about whales) a child might understand that the word *whale* refers to blue whales—but not whale sharks—or instead only know that a *whale* is a large animal that lives in the ocean. Children’s acquisition of the boundaries of visual categories and how they map onto words is a key developmental process, as children who lag behind their peers in vocabulary learning are at increased risk for language delays with significant academic and social consequences (Rescorla & Dale, 2013).

Most empirical work on the precision of children’s visual concepts has focused on the first few years of life, when children learn words at an astonishing rate (Bloom, 2000; Braginsky et al., 2019). In looking-while-listening tasks (Fernald et al., 2008), infants as young as 6 months look to events associated with common nouns like “bottle” and “banana” (Bergelson & Aslin, 2017), and 14- to 18-month-olds extend newly learned words to atypical exemplars of these categories (Weaver, Zettersten, & Saffran, 2024). By around their second birthday, children also generalize nouns to stylized, 3D exemplars (Pereira & Smith, 2009) as they learn that shape is a valuable cue to basic-level category membership. In parallel, empirical work on children’s developing ability to recognize objects (Ayzenberg & Behrmann, 2024) has focused on these first few years of childhood as an important period of developmental change.

An implicit assumption underlying much of this research is that preschoolers and school-aged children’s visual concepts—e.g., their knowledge of what whales “look like” —may not change substantially after they can understand the word “whale” in conversations. On this account, children associate words (e.g., “whale”) with high-fidelity visual information about a particular exemplar (or a few) during learning, and this is sufficient for acquiring the visual concept. Instead, the enrichment for this concept is thought to happen solely at the level of semantics (e.g., “whales eat krill”); on this *semantic enrichment hypothesis,* later learning during the elementary school years does not influence the content of the *visual* concepts per se.

But these assumptions may be incorrect. Perhaps these visual concepts themselves undergo considerable and gradual change beyond the first few years of childhood. On this alternative view, in parallel with learning an increasingly rich semantic space of word meanings, children also develop an increasingly detailed representation of the diagnostic features of these visual concepts, and a better understanding of the boundaries of these visual categories. Indeed, classic theories have long articulated a view of this type, in which children might initially set up a place-holder concept via “fast mapping” but then gradually fill in later details (Carey & Bartlett, 1978; cf. Swingley, 2010; Dale, 1965). An example from this early work focused on a case where three-year-olds remembered that a word (“chromium”) was a color word, but they did not know which color it was. Similarly, in foundational theories of vocabulary acquisition, children’s vocabulary knowledge has been thought of as progressing along stages (Dale, 1965; Beck et al., 1987), with an intermediate stage where children were thought to have a “vague knowledge of the word’s meaning” (Dale, 1965) before they fully grasp what a word means. For example, there might be an intermediate phase in which children understand that “whales” are animals that live in the ocean, but have less precision on the visual features that distinguish them from fish and other sea mammals.

Yet despite this long-standing recognition that children likely have partial word knowledge (Dale, 1965; Carey & Bartlett, 1978), all assessments of receptive vocabulary characterize word knowledge in an all-or-nothing fashion, perhaps because it is indeed difficult to quantify what it would mean to have a low-precision representation for a word (Dale, 1965; Carey & Bartlett, 1978). Both direct measures, such as the Peabody Picture Vocabulary Test (Dunn et al., 2003; Gershon et al., 2014), and parent-report measures of vocabulary knowledge (Braginsky et al., 2019) characterize children as either “knowing” a word or not. Since the distractors for these assessments are chosen carefully, children’s error patterns could, in principle, reveal evidence for partial knowledge (e.g., selecting a *whale shark* when the target word is *whale*), but the datasets resulting from these proprietary assessments are unavailable for re-analysis.

In parallel, a large literature on expertise suggests that visual learning almost certainly extends throughout early and middle childhood and into adulthood (for reviews, see Curby & Gauthier, 2005; Harel, 2016). Decades of empirical work has established that birding experts, car aficionados (Tanaka & Gauthier, 1997), and graphic artists (Perdreau & Cavanagh, 2013), among others, have both qualitatively and quantitatively different kinds of visual representations for the visual concepts that they have significant expertise with that is reflected in both behavioral and neural responses. Thus, on a *visual expertise hypothesis,* children’s visual concepts might continually be enriched throughout early and middle childhood. More recent developmental work supports this idea (Long et al, 2024): In a large observational study, children became increasingly able to both depict and recognize line drawings of visual concepts (e.g., “whale”, “clock”, “tiger”) from 3-10 years of age, hinting at underlying changes in how children represent the diagnostic features of these visual concepts.

Our theoretical accounts of visual concept development have thus been limited by (1) a need for larger developmental datasets with dense data on individual items and (2) the difficulty of systematically quantifying what it means to have a partial or imprecise representation of a visual concept. Here, we take steps to address both of these challenges. We created a new, open-source picture matching task and developed research partnerships with schools across the country in order to gather dense item data across a wide age range (3–14 years) from many children (*N*=3575 children, *N*=211 adults). We built a new, open-source “visual vocabulary” task where children heard a word (e.g., “swordfish”) and had to choose a picture in order to teach aliens about their meanings. We then collected data from a large sample of diverse children in both online contexts as well as through research-practice partnership with preschools and elementary school contexts across the U.S. Next, we leveraged modern computational models to choose distractor items systematically and to analyze children’s resulting responses. For each word we chose distractor items that varied in their linguistic similarity to the target word as operationalized via the similarity of the target and distractor words in the language encoder of a multimodal language model (Contrastive Language-Image Pre-training model, or CLIP; Radford et al., 2021), and we used embeddings from the visual encoder to model children’s error patterns. These advances in data and models thus allow us to quantify changes in the precision of children’s visual concepts across early and middle childhood for the first time.  
 With these data, we investigated three main questions. First, we examined whether we see evidence for continual learning of visual vocabulary knowledge throughout middle childhood (consistent with a visual expertise hypothesis) versus a plateau in performance in the elementary school years (consistent with a semantic enrichment hypothesis). Second, we examined evidence for changes in the precision of visual concepts before word knowledge is solidified, as per Dale’s (1965) account. If younger children's representations are initially overly broad, then we would expect younger children to select distractors of all types when they do not know a word’s exact meaning, and older children to almost exclusively choose high-similarity distractors (e.g., *whale sharks)*; if children bypass this stage where they have lossy representations, then children should choose distractors randomly. Finally, we examined children’s error patterns: if changes in children’s error patterns are driven by their knowledge of what visual information “goes with the word”, then a combination of linguistic and visual similarity metrics should explain variance in children’s error patterns. In contrast, if children are primarily drawn to visually-similar distractors when they do not know a word’s meaning—irrespective of their linguistic similarity to the target word—then visual similarity alone should be a driving factor in error patterns.

# Method

## Participants

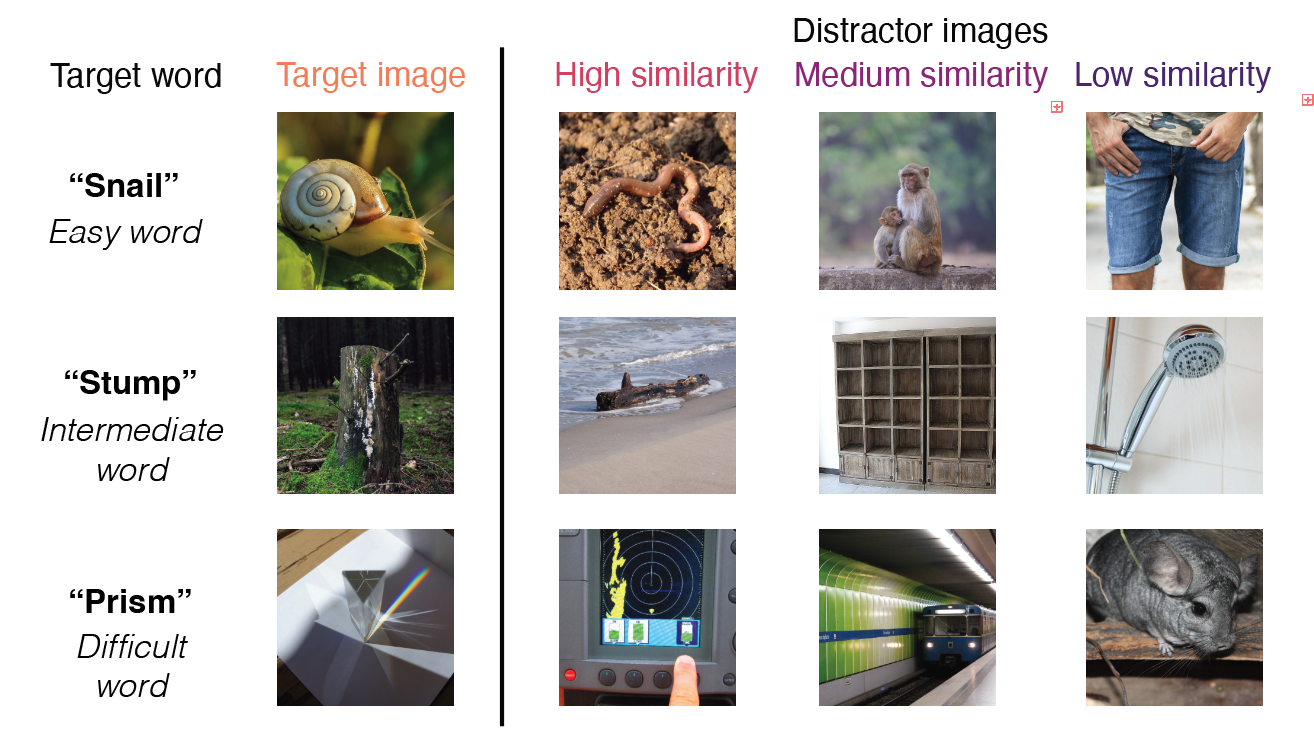
To obtain a large sample of responses for individual items across development, we collected data from children across several different testing contexts, totalling *N* = 3,786 unique participants. We collected data from children in an in-person preschool ( = 65, 3–5 year-olds), from the Children Helping Science (CHS) Platform, ( = 243, 3–7 year-olds), and representative samples from schools across the United States through the [BLINDED] platform ( = 3267, 5–14 year-olds) (BLINDED et al.). We also recruited adults online (= 211) via Prolific; we sampled from monolinguals and adults who spoke English as a second language to maximize our chance of finding item variability, and to measure a range of adult abilities. Schools included seven public or public charter school districts (14 schools), three independent schools focused on supporting students with learning differences, two independent schools, and one summer program; these diverse student data were obtained through research-practice partnerships. Most school-aged participants responded directly via clicking through the games on a laptop or tapping through the games on a tablet, except those recruited online through CHS and Prolific; children’s parents responded via clicking on the image that the child indicated on CHS, and adults responded via clicking on the images.

Children completed, on average, 25.02 trials that were sampled randomly from the stimulus set (maximum = 86; different maximum numbers of trials were included in different testing contexts). All participants who contributed data and scored above chance (chance = 25%, threshold of >30% accuracy) were included, regardless of the number of trials they completed (minimum trials = 2, maximum trials = 86).[[1]](#footnote-1) We thus excluded 84 participants who scored near chance on 4AFC trials (< 30% accurate) and were school-aged (>6 years of age); these participants completed an average of 17.72 trials.

## Materials

We capitalized on publicly available existing image and audio databases to generate trials. Visual concepts were taken from the THINGS+ dataset (Stoinski et al., 2023), after filtering out non-child safe images (e.g., weapons, cigarettes) and images with low-nameability (<.3), as per Stoinski et al., 2023. We used the copyright-free, high-quality image released for each visual concept. We subset to visual concepts that had available age-of-acquisition (AoA) ratings from a previous existing dataset (Kuperman, 2012) where adults were asked to retrospect on the age that they learned a word in development; these estimated AoAs were used as a proxy for item difficulty.[[2]](#footnote-2) Audio for each target word was taken from the MALD database (Tucker et al., 2019).   
 We used these stimuli to create trials that contained a target image paired with distractor images of similar difficulty (estimated using AoA) and varying similarity (estimated using a computational model). To do so, we sampled distractors with high, medium, and low similarity to the target word as operationalized via the embedding similarity of the words in the language encoder of a multimodal large language model (Radford et al., 2021).We determined high-, medium-, and low-similarity values relative to the distribution of all possible target-distractor similarity values for each word in the THINGS+ dataset. Stimuli were selected to optimize for having a maximum number of trials with unique target and distractors; in addition, we constrained the sampling such that target-distractor pairs had estimated AoA within 3 years of each other. For each target word, we first selected a high-similarity distractor that had the highest cosine similarity to the target (and was itself not one of the target words). For low-similarity words, we sampled a unique distractor word that had the lowest cosine similarity among the remaining distractors. For medium-similarity distractors, we then randomly sampled a distractor word that was the same animacy as the target word, and unique to the 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. All stimuli and their metadata are available on the public repository for this project (BLINDED).

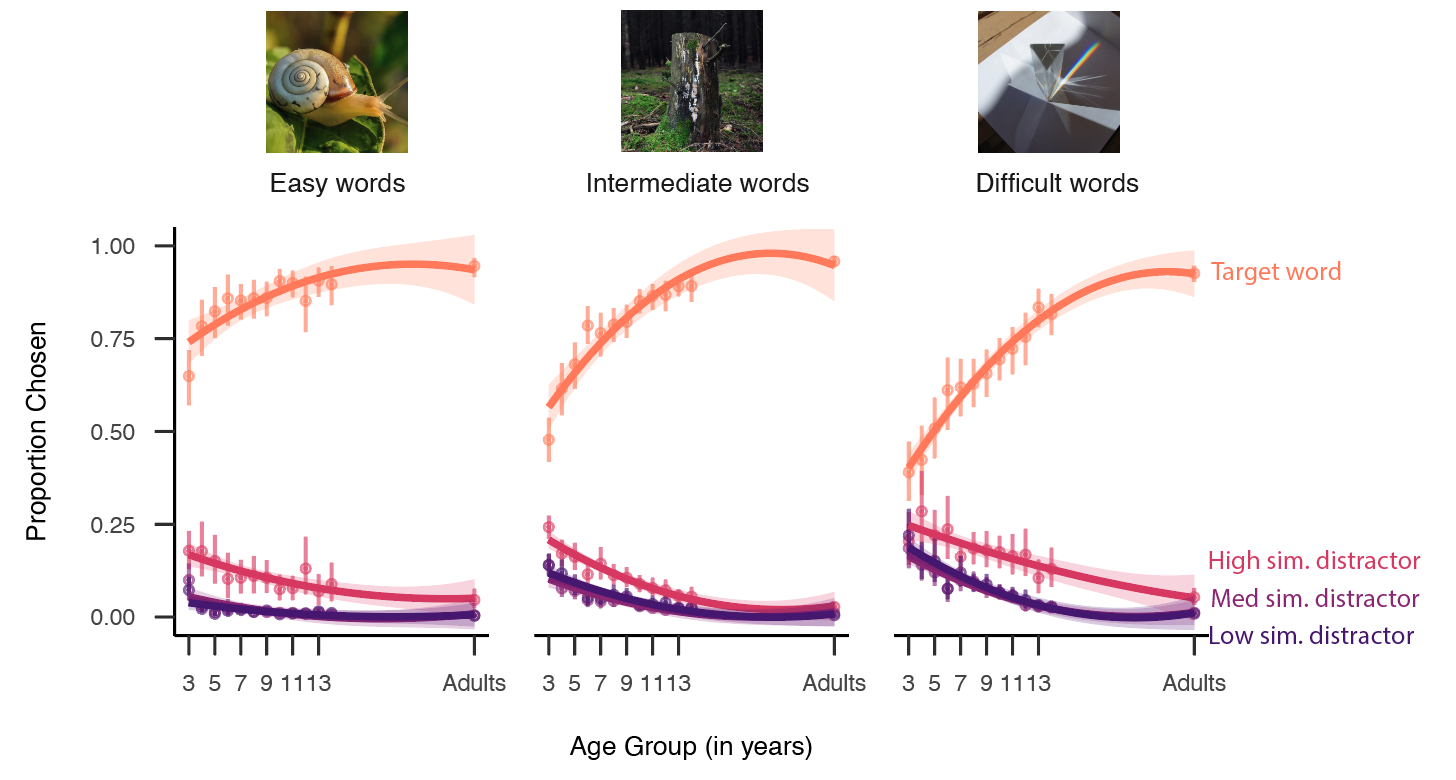
## Model features

We obtained all model features using the Open AI implementation of CLIP available at <https://github.com/openai/CLIP>. For language similarity, we computed the cosine similarity of the embeddings of the target word to each distractor word on each trial (e.g., *tulip – rose, tulip – glove, tulip – hubcap*). For visual similarity, we repeated this procedure but by obtaining image similarity vectors in the vision transformer for each target image and distractor image on each trial. 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. All code is available on the repository associated with this project.  
  
*Fig 1.* Example trials that depicted *easy*, *intermediate*, and *difficult* words, as operationalized via estimated age-of-acquisition (Kuperman et al., 2013); each target word image is shown alongside the three distractors that varied in similarity to the target word; distractors were chosen via the similarity of the target and distractor words in the language encoder of CLIP (Radford et al., 2021).

**Procedure**

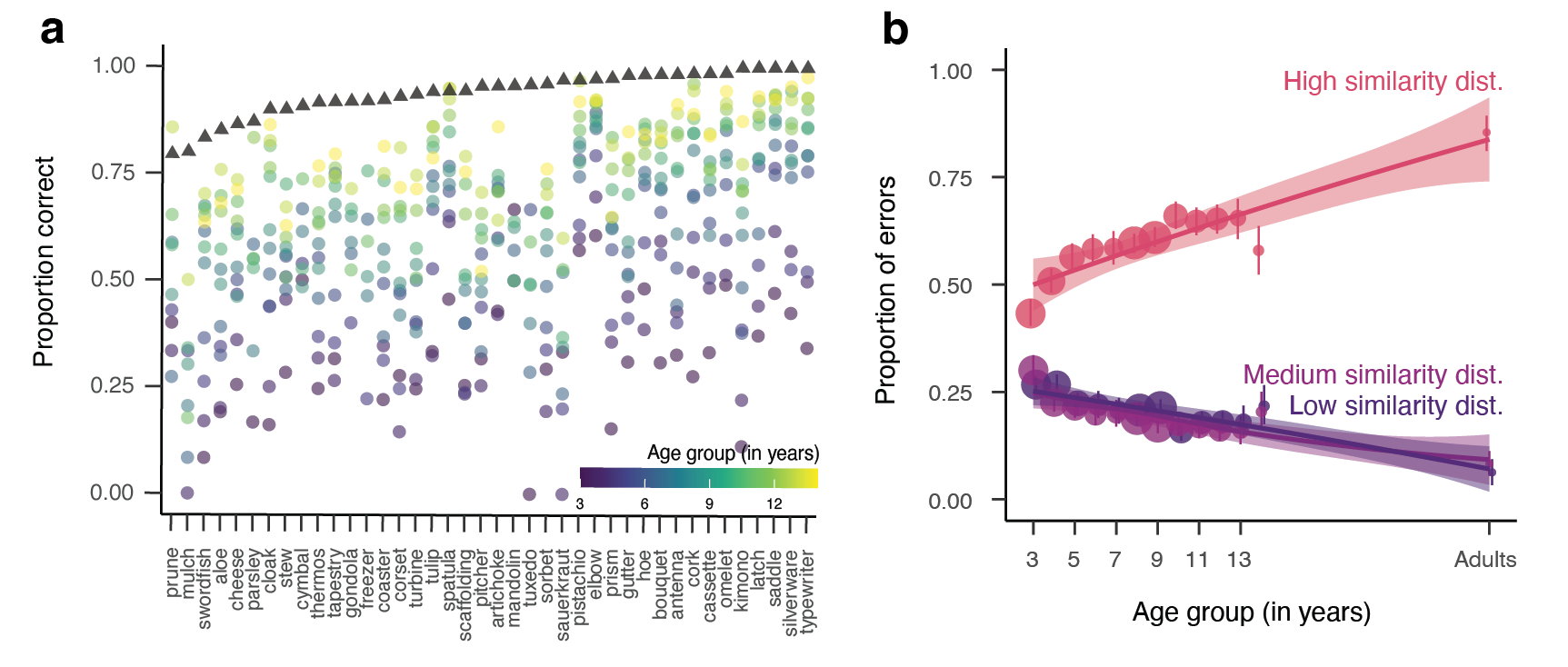
Children were invited to participate in a picture matching game where they were asked to “teach aliens on another planet about some of the words on *our* planet”; children picked a particular alien to “accompany them on their journey.” Before the images appeared, children heard a target word (e.g., “apple”) 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 (see example trial in Figure 1b). On practice trials, the distractor images were all very dissimilar to the target concept, and the target word was relatively easy. The game played a chime sound if children chose the correct image, and a slightly unpleasant sound if they responded incorrectly; this feedback ensured that children of all ages understood the task correctly. 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 whenever they wanted to. Different versions of the game included varying amounts of total possible trials and additional items, as these games were developed as part of an additional project to create an open-sourced measure of children’s vocabulary knowledge (Blinded et al., in prep). Here, we analyze children’s responses to items that were generated using the THINGS+ dataset with distractors of varying difficulty (see Stimuli).

# Results

  
*Fig 2.* Visual vocabulary task performance as a function of the age of the child completing the task, plotted separately for relatively easy, intermediate, or difficult words; words are binned into terciles based on the estimated AoA from Kuperman et al., 2012. Lines refer to the proportion of words that children chose the target (orange), high-similarity (pink), medium similarity (darker pink), or low similarity (purple) distractor at each age; error bars represent bootstrapped confidence intervals. Example images depict three target concepts (*snail, stump, prism)* shown in Figure 1.

## A protracted developmental trajectory

Do children gradually acquire visual vocabulary knowledge, or are precise representations for most nouns acquired by middle childhood? We found a gradual increase across our age groups in how well children could correctly identify the target picture; Figure 2 shows the proportion of time that children identified the target picture, highlighting a protracted developmental trajectory and little evidence for a plateau in the elementary years. Older children were more accurate at identifying the correct referent that a word referred to, with the oldest children in our sample (13- and 14-year-olds) still performing less accurately than adults.   
 We found a slight developmental trend for relatively easy words that had an average estimated age-of-acquisition (Kuperman et al., 2012) (AoA) of 4.81 years (*SD* = 0.87; *acorn, bulldozer, scoop, puddle, sprinkler)*, but change over development was much more pronounced for more difficult words (*barrel, buffet, coaster, stump, carousel;* average AoA = 6.95 years, = 0.65) and challenging words (e.g., *tapestry, scaffolding, mandolin, aloe;* average AoA = 9.60 years, = 1.21) (see Figure 2). At an item level, 3-year-olds’ performance on the easiest words (*watermelon*, *M* = .932*; turtle, M* = .933, *lollipop*, *M* = .932) suggests that they understood the task. However, other items with relatively low AoAs (*squirrel*, *M* = .768, *hamster*, *M* = .702; *map*, *M* = .709) were more difficult for young children.   
 Over development, the words that showed the greatest change across age (see Figure 3a), included some animals (e.g., *swordfish*), as well as inanimate objects (*prism, antenna, sandbag, turbine*), but also other kinds of objects, including parts of buildings (*scaffolding, gutter*). Some items, however, showed relatively little change across age: these tended to be easy words that children across all ages identified correctly (*sunflower*, *rice, potato, hedgehog)*. Overall, these results highlight variability across individual items and suggest that children gradually build knowledge of many visual concepts throughout childhood.

*Fig 3.* (a) Visualization of top 30 items that showed the largest differences across age groups; values represent the proportion of correct responses on each target word (e.g., *swordfish*) within each age group. Triangles represent data from both English monolinguals and second-language learners collected via Prolific. Items are ordered by difficulty for adults. (b) Changes in the proportion of errors chosen as a function of children's age, where green lines reflected higher similarity distractors. Dot size represents the number of errors made by children in each age group. Error bars represent 95 percent bootstrapped confidence intervals.

## Changes in the precision of visual concepts

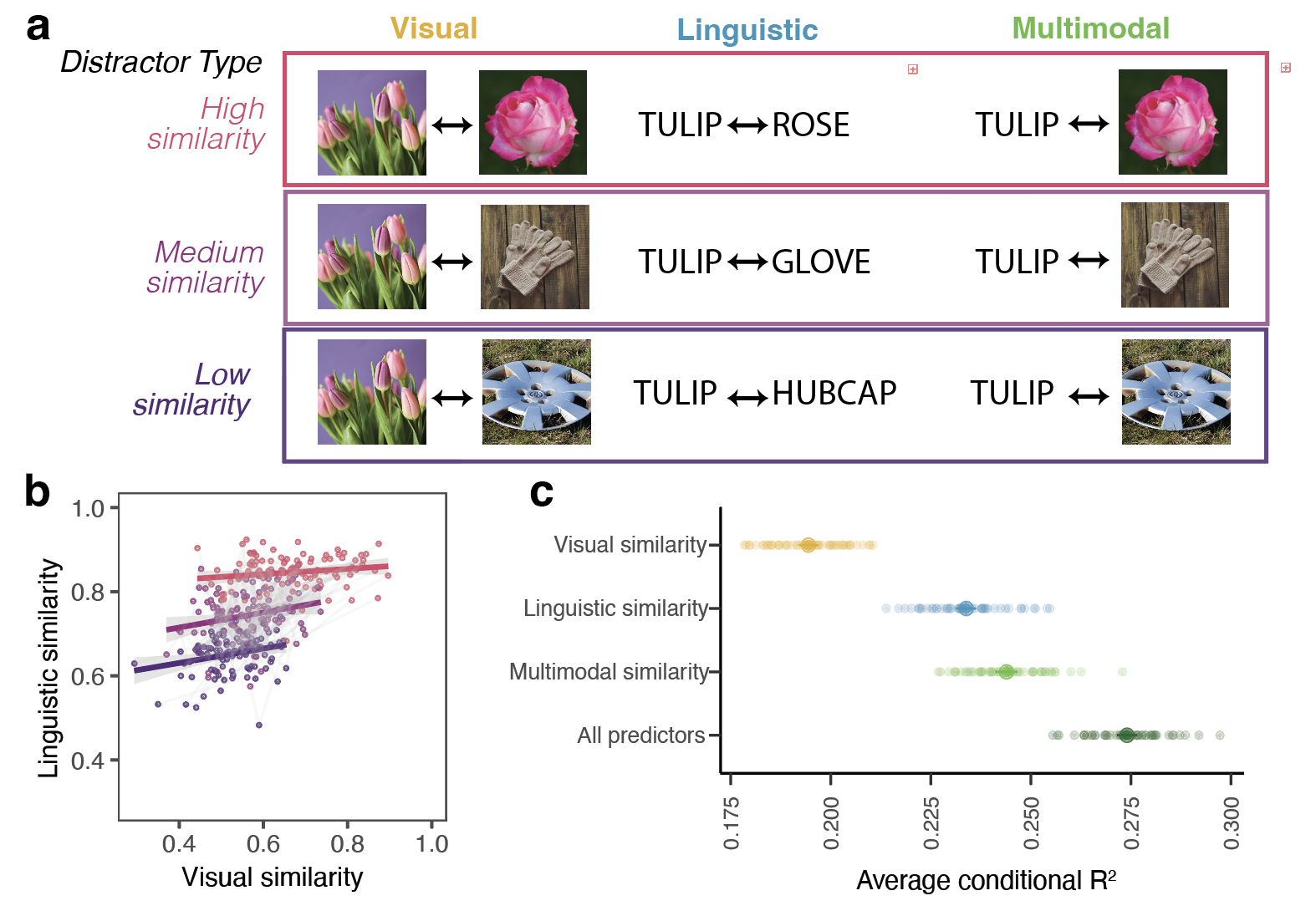
Do children begin with high-fidelity representations of what word referents look like or are they acquired more slowly? Thus far, these analyses support a view where children progressively acquire visual vocabulary knowledge via precise visual concepts (e.g., that can distinguish between a *whale* and a *whale shark*) on a relatively long developmental timeline. However, they do not yet establish whether children have visual concepts that become more precise en route to mature word understanding. If children's visual concepts indeed have an intermediate stage where they are overly broad (e.g., encompass both *whales* and *whale sharks*), then we would expect younger children to be more likely to choose distractors of all types (showing evidence for a very coarse representation or no representation at all), whereas we would expect older children to almost exclusively choose related distractors (showing evidence of partial knowledge). Alternatively, if children quickly acquire relatively high-fidelity representations, then we should only observe changes in how accurate children are at identifying the target word, with no changes in the types of distractors that children choose when they choose incorrectly.

We examined whether there are systematic changes in how children made errors across development, shown in Figure 3b. Consistent with the partial knowledge hypothesis, we found that older children chose more closely related distractors than younger children. Adults were still more likely to choose the related distractors than the oldest children in our sample (14-year-olds). We confirmed this result via a linear regression, modeling the proportion of errors that each child chose related distractors as our dependent variable as a function of children’s age (in years); we also included the number of errors each child made as a covariate as this varied widely by participant and age group. We found a main effect of age (*B =* 0.04, *SE* = 0.01, *t* = 7.45, *p* < .001): older children were more likely to choose linguistically related vs. unrelated distractors.[[3]](#footnote-3) However, we did not find strong differences between distractors that were less related to the target word: both low-similarity and medium-similarity distractors were equally unlikely to be chosen, suggesting that even though model similarities distinguished these options, children did not. Overall, these results suggest a progression where children gradually build up detailed knowledge about the visual referents of many challenging words.

## Modeling changes in children’s error patterns

In a final set of analyses, we aimed to understand the source of changes in children’s error patterns by leveraging similarity metrics derived from a large, multimodal language model (Radford et al., 2021). When we created our stimuli, we paired targets and distractors by using word similarity in a linguistic embedding space, but of course the corresponding visual stimuli naturally vary in their visual similarity to each other. This is true both at a category level— that is, animals are more similar visually to each other than to inanimate objects (Long et al., 2017; Krigeskorte et al., 2008) but also can be true at an image level—e.g., the flowers in the images of *rose* and *tulip* were both pink. Thus, some stimuli on certain trials could be more related to the target concept semantically and less related visually, or vice versa. For example, the words *scoop* and *sauce* had high similarity in the language encoder (*r* = .924), and their corresponding images had lower similarity in the vision encoder (*r* = .560). Conversely, the words *tulip* and *rose* had relatively lower similarity in the language encoder (*r* = .785) than their corresponding images in the vision encoder (*r* = .872) (see Figure 4A, Figures 4B). Thus, our large and varied set of stimuli—and the ability to easily parameterize the relative linguistic and visual similarity of all target-distractor pairs—affords the opportunity to examine what drives item-level variability in this task. We thus examined the degree to which children’s error patterns reflected changes in how they processed the visual similarity of the targets and distractors, their semantic similarity, or—perhaps most likely—some combination.

To do so, we used a series of cross-validated linear mixed effect models, where we examined the degree to which visual, linguistic, and multimodal similarity metrics (and their combination) derived from multimodal large language model embeddings could explain children’s error patterns. Specifically, we modeled the proportion of time that participants chose each distractor for a target word as a function of the age (in years) of the participants, and (1) the similarity of the target word to each distractor word (linguistic embeddings), (2) the similarity of the target image to each distractor image (visual embeddings), (3) the similarity of the target word to each distractor image (multimodal embeddings), and (4) a model with all predictors combined; we additionally added the total number of errors per item as a covariate, with random intercepts for each target word. We iteratively sampled 80 percent of the dataset 50 times, and then evaluated the conditional R-squared for each model for each split; these values are plotted in Figure 4C. These results revealed that combining both the visual and linguistic embeddings—either together in one, large mixed-effect model, or via multimodal embeddings—led to increased explained variance in error patterns. These results thus suggest that changes in error patterns across age are not solely due to changes in children’s ability to reject the distractor images that are visually similar to the target concept. Rather, children’s errors are driven by a combination of both visual and linguistic similarity.

***Figure 4.*  (a) Schematic of the three different ways that embedding similarity was calculated in CLIP (Radford et al., 2021) as a function of the different distractor types. (b) Visualization of the degree to which visual and linguistic similarity are collinear in this stimulus space, with each dot representing similarity calculated by a target-distractor pair; colors represent the distractor type depicted in (a). (c)Average explained variance in children’s error patterns (conditional R-squared across 50 linear mixed effect models with subsamples of the dataset, see *Methods*). This is shown for models with only linguistic, visual, and multimodal predictors, and for a combined model with all predictors. Error bars represent bootstrapped 95 percent confidence intervals across 50 model iterations.

# Discussion

Do children acquire visual concepts in an all-or-nothing fashion or gradually acquire partial knowledge? We collected a large dataset of picture-matching performance in children 3–14 years of age where the distractors were systematically related to the target concept. Children from older vs. younger age groups were more accurate at identifying the referents of words; in addition, children’s error patterns progressed from relatively random towards systematically choosing similar distractors. Models incorporating both the visual features of the images and the semantic similarity of the target and distractor words best explained children’s error patterns. These analyses suggest a gradual transition from coarse to finer-grained visual and linguistic representations over early and middle childhood, with abundant partial knowledge along the way.

These data suggest that there is substantial enrichment in children’s existing representations of visual concepts throughout childhood. On this account, even school-aged children’s visual concepts may undergo substantial change as they learn more about the world around them and the structure of the categories in it. For example, certain visual features may become more salient in children’s representations as children understand their functional roles (e.g., *camels* have humps to store water) or the degree to which they delineate a category boundary (e.g., between *whales* and *whale sharks*). Even after a child can successfully understand a word in a supportive context, their underlying visual concept may still change throughout later childhood. In addition, while adults likely have precise representations for visual concepts they interact with frequently—such as *trees, cups,* and *cars—*they likely cannot distinguish between many categories unless they are relative experts in that particular domain, for example, *gazelles* vs. *antelopes*.

These findings raise new questions about how exactly children’s visual concepts are changing. For example, if a child picks a related distractor (e.g., *whale shark*) when they hear a target word (*whale*), we cannot tell exactly why from these data. Children’s visual concepts could be both too narrow—excluding the particular exemplar of the whale that represents the target concept—and also too broad at the same time, including the “whale shark” as part of the “whale” concept. Future studies that iteratively test the boundaries of individual children’s visual concepts could help delineate the evolving content of their visual concepts, how they vary across individuals, and how they change across development (Leon Villagra et al., 2022).

If visual concept learning extends throughout development, more learning mechanisms and environments are relevant. Learning environments extend beyond the home and into educational contexts; children’s learning partners include their peers, teachers, and siblings, and children’s individualized interests may influence how they acquire detailed visual representations for words. In these contexts, children likely learn about visual features from generic utterances (e.g., “*Tigers* have stripes”) (Gelman, Ware, & Kleinberg, 2010). Semantic knowledge (e.g., “*whales* are mammals”) also likely influences which visual features are most prominent in children’s representation. Our theories of visual concept learning, accordingly, need to expand beyond thinking about how children generalize from labelled groups of exemplars.

There are several limitations to the current work. While we include data from a diverse group of children over a wide developmental age range, our conclusions are drawn from just over one hundred experimental stimuli; expanding the range and diversity of both the participants and the visual concepts in our study (Henrich, Heine, & Norenzayan, 2010) is necessary to understand the generalizability of these findings. Further, our data are cross-sectional, and thus cannot provide evidence for changes in the precision of representations within individual minds. Longitudinal data could confirm the theories suggested here.   
 The present work highlights the promise of large-scale, online platforms for collecting datasets to examine the development of visual concepts. Our findings provide evidence that children have abundant partial knowledge for many words before they learn their precise meanings, and highlight that children’s visual concepts change gradually throughout middle childhood. More broadly, these results expand the timeline and mechanisms relevant for visual concept learning.

# Acknowledgments

This was supported by a [BLINDED] to [BLINDED]. We gratefully acknowledge the children, families, and schools who participated in these studies, and members of the [BLINDED] and [BLINDED] labs who provided valuable feedback on the design of these studies. We also thank [BLINDED] for aiding with initial stimuli and task design.

# References

Ayzenberg, V., & Behrmann, M. (2024). Development of visual object recognition. *Nature Reviews Psychology*, *3*(2), 73–90. <https://doi.org/10.1038/s44159-023-00266-w>

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Bloom, P. (2000). *How children learn the meanings of words*. Boston, MA: MIT Press.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind*, *3*, 52–67. https://doi.org/<https://doi.org/10.1162/opmi_a_00026>

Curby, K. M., & Gauthier, I. (2010). To the trained eye: Perceptual expertise alters visual processing. Topics in Cognitive Science, 2(2), 189-201.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English,* 42(8), 895–901.

Dunn, L. M., Dunn, D. M., & Bulheller, S. (2003). *Peabody picture vocabulary test: PPVT*. Swets Test Services.

Fernald, A. E., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In *Developmental psycholinguistics: On-line methods in children’s language processing* (pp. 97–135). John Benjamins Publishing Company.

Gelman, S. A., Ware, E. A., & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive Psychology*, *61*(3), 273–301.

Gershon, R. C., Cook, K. F., Mungas, D., Manly, J. J., Slotkin, J., Beaumont, J. L., & Weintraub, S. (2014). Language measures of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *20*(6), 642–651.

Harel, A. (2016). What is special about expertise? Visual expertise reveals the interactive nature of real-world object recognition. *Neuropsychologia*, 83, 88-99.

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

Leon Villagra, P., Ehrlich, I., Lucas, C., & Buchsbaum, D. (2022). Uncovering children’s concepts and conceptual change. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 44, No. 44).

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children’s production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191.

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

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.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.others. (2021b). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.others. (2021a). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763. PmLR.

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

Suggate, S., Schaughency, E., McAnally, H., & Reese, E. (2018). From infancy to adolescence: The longitudinal links between vocabulary, early literacy skills, oral narrative, and reading comprehension. *Cognitive Development*, *47*, 82–95.

Tanaka, J., & Gauthier, I. (1997). Expertise in object and face recognition. *Psychology of Learning and Motivation*, *36*, 85–126.

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.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263.

Weaver, H., Zettersten, M., & Saffran, J. R. (2024). Becoming word meaning experts: Infants’ processing of familiar words in the context of typical and atypical exemplars. *Child Development*, *95*(5), e352–e372.

1. Some participants participated in the game more than once (*N* = 214 children), and their data are included as if they completed additional trials; these were often additional items due to random sampling. Note that our statistical models include random intercepts for each participant and operate at the trial level; however, excluding these participants did not change the pattern of results [↑](#footnote-ref-1)
2. Item data from a parental report of vocabulary knowledge (Frank et al., 2021) was not available for the majority of the words in the dataset due to their difficulty. [↑](#footnote-ref-2)
3. We also modeled these effects at the item level in a linear mixed-effect model, including the estimated age-of-acquisition for each word as a covariate and with random intercepts for each item, finding a fixed effect of age, and thus the same pattern of effects. All analyses are available on the public repository for this project. [↑](#footnote-ref-3)