All Creatures Great and Small: Category-Relevant Statistical Regularities in Children's Books

Layla Unger (unger.114@osu.edu)
Ohio State University, Department of Psychology, 1835 Neil Avenue
Columbus, OH 43210 USA

Anna V. Fisher (Fisher49@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15214 USA

Abstract

Sensitivity to statistical co-occurrence regularities is present from infancy. This sensitivity may contribute to learning in many domains, including category learning. However, prior research has not examined whether everyday input conveys category-relevant statistical regularities. This study assessed whether statistical regularities relevant to real-world categories are present in a commonly experienced source input – children's picture books. We focused on animal categories because this is a domain in which children receive much exposure from an early age, while simultaneously holding persistent misconceptions about category membership beyond preschool years. Analysis of 80 books revealed that they: 1) Were likely to contain regularities from which individual species categories (e.g., "chicken") might be learned, but 2) Were unlikely to contain regularities from which broader taxonomic categories (e.g., "bird") might be learned. These findings point to a paucity of taxonomically-relevant statistical regularities that may contribute to persistent taxonomic misconceptions.

Keywords: Cognitive Development; Semantic Knowledge; Semantic Development; Category learning

Introduction

Over the course of development, we are faced with the challenge of acquiring an extensive body of knowledge about the world. For instance, we must learn to classify the entities we perceive around us into meaningful categories, segment continuous visual input into events, and acquire our ambient language. Research in the field of cognitive development over recent decades has highlighted how these feats of knowledge acquisition may be facilitated in part by a sensitivity to statistical regularities in the environment. In some domains, this research has included investigations into the nature of the statistical regularities that are available in the environment for learners to pick up on. For example, in the domain of language acquisition, studies have investigated the degree to which spoken and written language input contain statistical regularities to which learners are sensitive, such as the reliable co-occurrence of speech sounds (Mattys & Jusczyk, 2001; Pelucchi, Hay, & Saffran, 2009) In contrast, in other domains where sensitivity to statistical regularities may play a critical role in knowledge acquisition, such as category learning, we know little about the statistical regularities present in the environment.

The overall aim of the present study was to investigate the degree to which statistical regularities that are relevant to real-world category learning are actually present in the environmental input that learners receive, starting early in development. To conduct this study, we made a set of evidence-based choices about which real-world categories, statistical regularities, and sources of environmental input to investigate.

First, several factors prompted us to focus on statistical regularities that may foster knowledge of taxonomic categories of animals, such as "bird", "mammal", and "reptile". We focused on taxonomic categories because they are cognitively useful for processes such as inductive reasoning and integrating new information into memory. The utility in inductive reasoning comes from the fact that membership in a taxonomic category tends to reliably predict the properties that an entity possesses (e.g., if an animal is a mammal, it will have a four-chambered heart), and therefore facilitates inferences about such properties without the need for direct observation (e.g., Heit, 2000). Moreover, recent evidence suggests that taxonomic category knowledge may facilitate the integration of newly-learned information about members of the category into memory (Pinkham, Kaefer, & Neuman, 2014).

We specifically focused on taxonomic categories the animal domain in part because it is one in which environmental input is common and familiar from an early age. Moreover, the possibility that statistical regularities can shape category knowledge in the animal domain is supported by evidence that children show earlier and stronger knowledge that animals from the same taxonomic category are related to each other when those animals also co-occur, versus when they do not (Unger, Fisher, Nugent, Ventura, & MacLellan, 2016). However, evidence from other lines of research suggest that the statistical regularities children experience in the animal domain may often instead lead them to form an inaccurate understanding of taxonomic category composition. Specifically, findings from multiple learning sciences studies suggest that children possess persistent misconceptions about taxonomic category membership in the domain of animals that is suggestive of the contributions of learning from statistical regularities. In comparison to children's robust knowledge of the habitat groups within which animals regularly co-occur (e.g., farm animals, water

animals, etc.) (Crowe & Prescott, 2003; Kattmann, 1998, 2001; Storm, 1980), their knowledge of taxonomic category membership (e.g., mammal, bird, reptile, etc.) is often poor. For example, in a large-scale study of over 450 U.S. students from elementary school, middle school, high school, and college, Trowbridge and Mintzes (1988) found that substantial numbers of students throughout the age-range miscategorized many animals, such as identifying all animals that live in water as fish, including whales and marine invertebrates. Such misconceptions persisted even in college students who majored in biology (e.g., 20% of biology majors identified whales as fish). Such misconceptions appear cross-cultural (Yen, Yao, & Chiu, 2004). Taken together, these findings suggest that children form categories of animals in part based on the regularities with which different animals co-occur, and that such categories are often inconsistent with taxonomic category membership. By the same token, these findings suggest that the statistical regularities that children experience in the animal domain are not informative about the composition of taxonomic category membership.

Second, we chose as our statistical regularity of interest the regularity with which a set of perceptual inputs co-occur more reliably with each other than with others. Throughout the remainder of this paper, we refer to these as "co-occurrence regularities". This choice was motivated in part by prior research that has shown that a sensitivity to this regularity is present from very early in development (possibly birth, e.g., Bulf, Johnson, & Valenza, 2011). Additional motivation for this choice came from cognitive neuroscience, computational modeling and behavioral evidence that co-occurrence may foster category knowledge by increasing the similarity of mental representations of co-occurring entities (Huebner & Willits, 2017; Schapiro, Kustner, & Turk-Browne, 2012; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013).

Finally, the source of environmental input in which we investigated animal taxonomic category-relevant co-occurrence regularities was children's picture books. One motivation for this choice was evidence that picture books represent a common source of environmental input from very early in development (e.g., Young, Davis, Schoen, & Parker, 1998) This choice was also motivated by evidence that children's picture books are a key source from which children learn about animals (Ganea, Ma, & DeLoache, 2011). Lastly, we focused on children's books because they readily lend themselves to quantifying co-occurrence regularities by recording which items co-occur on the same page.

Present Study

For the reasons outlined above, we investigated the degree to which category-relevant statistical regularities are present in environmental input by measuring the degree to which animals from the same taxonomic category are likely co-occur on the same page in a large sample of children's picture books. Moreover, we analyzed the degree to which such co-occurrences are likely to be informative about the composition and breadth of taxonomic categories. Specifically, evidence from prior category learning research (Oakes, Coppage, & Dingel, 1997; Perry, Samuelson, Malloy, & Schiffer, 2010) suggests that examples of many different items from the same category are more informative about the composition and breadth of that category than are multiple examples of the same item. Therefore, we tested whether co-occurrences between members of the same taxonomic category are primarily between animals of the same species (e.g., multiple cows depicted on the same page), or are also between diverse members of the category (e.g., multiple different kinds of mammal depicted on the same page). Finally, to investigate whether the nature of animal category-relevant statistical regularities to which learners are exposed changes with age, we measured the relationship between the regularities present in children's books and the target age range for which each book was written.

Methods

Book Sample

The sample of books analyzed consisted of 80 children's books about or featuring animals. To afford an evaluation of co-occurrences between depicted animals, only books in which multiple pages depicted more than one animal were included in the sample. To ensure that books were representative of those to which children are typically exposed, books were selected on the basis of librarian recommendations, circulation statistics for children's books from the local public library, and amazon.com best sellers.

The selection of books was additionally guided by a consideration of the diversity of books to which children may be exposed. Specifically, books were selected to assemble a sample that included both fiction and non-fiction books written for a wide range of target age groups (early childhood through middle school-age). However, due to the real-world distribution of picture books about animals across genres and target age groups, the sample was biased towards non-fiction books written for pre- to early-school-age children. Both the genre and target age group of each book was recorded for use in analyses, as described in the Results and Discussion section below.

Book Coding Schemes

The books in the sample were coded by hypothesis-blind research assistants in order to yield data from which rates of co-occurrences of taxonomically related versus unrelated animals could be measured. The coding procedure was therefore designed to record the taxonomic category of each animal that appeared on each page of each book. The taxonomic categories identified were: Mammal, Bird, Fish, Reptile, Amphibian, Insect, Arachnid, Crustacean, Mollusk, and Other Invertebrate. Additionally, pairs of contiguous pages that are visible at the same time were treated as a single "page", because the animals they depict would be experienced as co-occurring. When large numbers of animals (e.g., more than 50) were depicted on a page, coders were instructed to estimate the total number by counting the

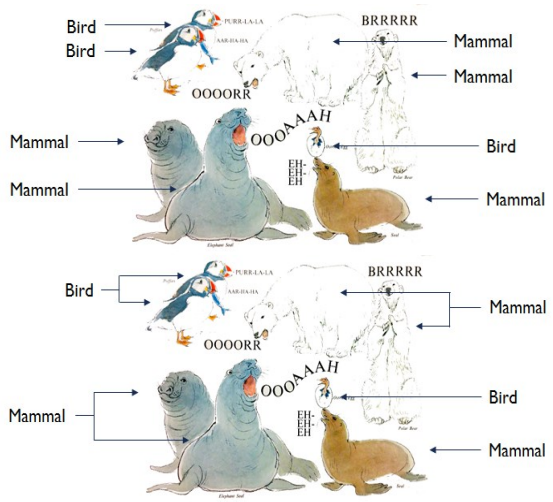

Figure 1. Example of the taxonomic categories of animals that would be recorded on a page using the All Animals scheme (top) and the Animal Groups scheme (bottom).

number within a square inch of the page, and multiplying this number by the number of square inches that comprised the area over which the animals were depicted.

This general approach was used to code the contents of books in the sample according to two coding schemes. Two coding schemes were developed in order to take into account the possibility that co-occurrences between a diverse set of animals from the same taxonomic category may be more informative about the composition of the category than co-occurrences between very similar animals. For example, a learner may acquire a more complete and accurate sense of what animals are mammals when presented with a variety of mammals, rather than multiple exemplars of the same mammal. This possibility is consistent with evidence that children learn more inclusive categories when presented with heterogeneous versus homogeneous exemplars (Oakes et al., 1997; Perry et al., 2010).

The two coding schemes therefore consisted of: 1) An "All Animals" scheme, in which the taxonomic category of each individual animal on each page was recorded, and 2) An "Animal Groups" scheme, in which multiple exemplars of the same animal on a page were all grouped as a single instance of their respective taxonomic category (Figure 1). For example, if a page depicted three cats, it would be coded as depicting three instances of the Mammal category according to the All Animals scheme, but only one instance according to the Animal Groups scheme. However, if a page depicted three different mammals such as a cat, moose, and rabbit, it would be coded as depicting three instances of the Mammal category according to both schemes. The degree to which the different coding schemes yield different descriptions of the rate at which children experience co-occurrences between animals from the same taxonomic category is evaluated in the Results and Discussion section below.

Results

Data collected via the All Animals and Animal Groups coding schemes were processed separately following the same approach. First, for each page, the number of co-occurrences between pairs of animals from the same and from different taxonomic categories was calculated. For example, if a page was coded as depicting three instances of the Mammal category and two instance of the Fish category, the number of same taxonomic category co-occurrences would be recorded as four, and the number of different taxonomic category co-occurrences would be recorded as six[1]. Next, the total number of same and different taxonomic category co-occurrences was calculated for each book. Finally, the proportion of same taxonomic category co-occurrences out of total number of co-occurrences was then calculated for each book in order to assess the rate at which children are exposed to taxonomic co-occurrence regularities.

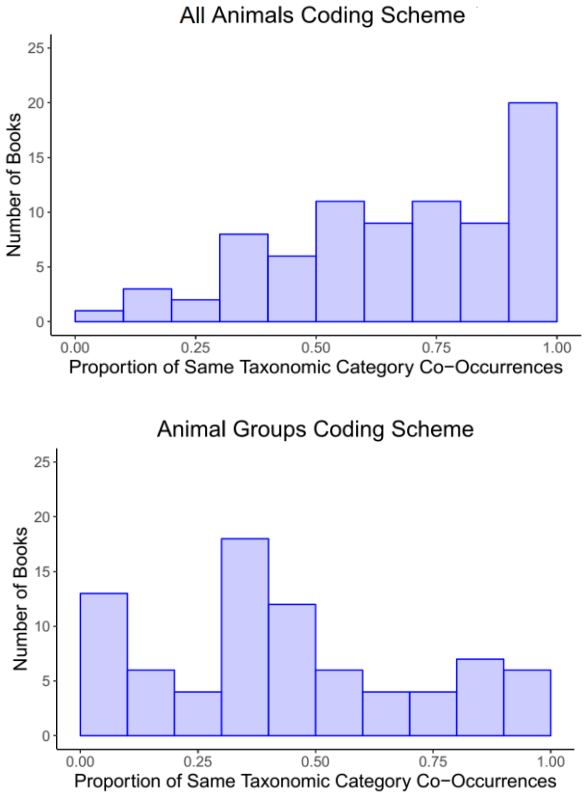

Figure 2. Distributions of proportions of same taxonomic category co-occurrences across books according to the Individual and All Animals coding schemes.

[1] For the purpose of illustration: Consider the three mammals as M_1, M_2, and M_3, and the two fish as F_1 and F_2. The same category co-occurrences would be M_1-M_2, M_1-M_3, M_2-M_3, and F_1-F_2. The different category co-occurrences would be M_1-F_1, M_1-F_2, M_2-F_1, M_2-F_2, M_3-F_1 and M_3-F_2.

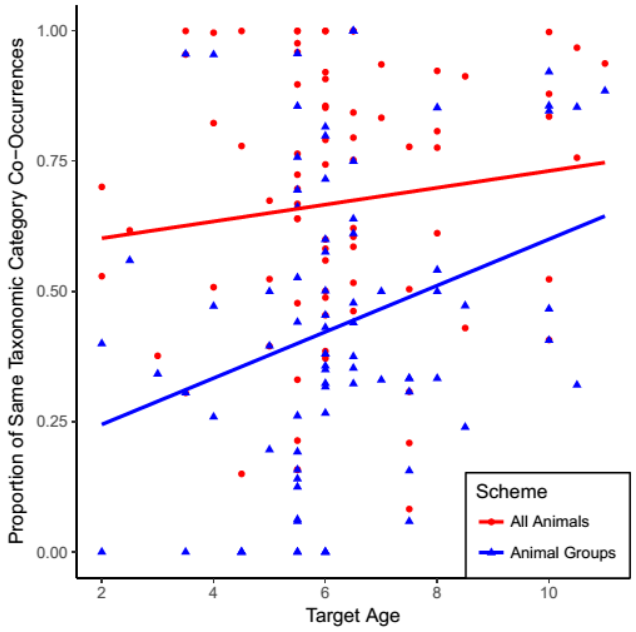

Figure 3. Scatterplot with best-fit lines showing increase in rates of same taxonomic category co-occurrences with age in data collected using the All animals and Animal Groups coding schemes.

The distributions of same taxonomic category co-occurrence proportions across books in the sample according to the All Animals and Animal Groups coding schemes is depicted in Figure 2. Examination of these distributions reveals that the shape of the distribution appears influenced by the coding scheme. According to the All Animals coding scheme, books with high proportions of same taxonomic category co-occurrences are relatively common, whereas according to the Animal Groups coding scheme, they are uncommon. These patterns indicate that the high rates of same taxonomic category co-occurrences in books at the upper end of the distribution for the All Animals coding scheme data are likely to have been recorded due to the depiction of large numbers of the same animal (e.g., schools of fish, colonies of ants, etc.), rather than multiple different exemplars of a taxonomic category. When such depictions are recorded as one instance of a given category in the Animal Groups coding scheme, rates of same taxonomic category co-occurrence appear much lower.

Predictors of Same Taxonomic Category Co-Occurrence

To quantitatively evaluate the influence of both coding scheme and book characteristics, including target age group and genre (fiction versus non-fiction), we followed a multi-level mixed effects modeling approach in which we first predicted the outcome variable of Same taxonomic co-occurrence proportion using a null model that included one random effect of Book, then tested whether the addition of Coding Scheme (0=All Animals, 1=Animal Groups), Genre (0=Fiction, 1=Non-Fiction), and Target age group as fixed effects improved the model fit. For the purpose of this analysis, the Target age group for a given book was taken as the age at the midpoint of the range of target ages for that book. For example, the Target age group a book written for children aged 3-5 years was recorded as 4.

The results of this analysis revealed that, as inspection of the distributions for the two coding schemes in Figure 2 suggests, the addition of Coding Scheme to the null model improved model fit ($\chi^2(3,4)=40.27$, $p<.001$). The additional inclusion of Target age group further improved model fit ($\chi^2(4,5)=5.19$, $p<.05$). However, model fit was not improved by either the inclusion of an interaction between Coding Scheme and Target age ($\chi^2(5,6)=2.67$, $p=.10$), or the inclusion of Genre ($\chi^2(5,6)=0.05$, $p=.83$). The parameters of the null and two models that incrementally improved fit are reported in Table 1. The estimates for the fixed effect of Coding Scheme indicate that, in comparison to the All Animals scheme, the Animal Groups scheme yielded lower proportions of same taxonomic category co-occurrences (which is consistent with distributions depicted in Figure 2). Moreover, the estimates for the fixed effect of Target age

Table 1.
Estimates of effects for Proportion of same taxonomic category co-occurrences

	Null Model			Coding Scheme Model			Coding Scheme+Target Age Model		
	Estimate	*SE*	*t*	*Estimate*	*SE*	*t*	*Estimate*	*SE*	*t*
Fixed Effects									
(Intercept)	0.55	.03	21.54	0.67	0.03	21.99***	0.48	0.09	5.49***
Coding Scheme				-0.24	0.03	7.19***	-0.24	0.03	7.19***
Target Age							0.03	0.01	2.89*
	Variance		*SD*	*Variance*		*SD*	*Variance*		*SD*
Random Effects									
Book	0.02		0.13	0.03		0.17	0.03		0.17

Note. *$p<.05$, **$p<.01$, ***$p<.001$

indicate that proportions of same taxonomic category co-occurrences increased with the age group for which a book was written (see Figure 3 for graphical depiction of these patterns).

Discussion

Knowledge of taxonomic categories categories is cognitively useful for processes such as inductive inference and integrating newly learned information into memory (Heit, 2000; Pinkham et al., 2014). However, even in a domain that is commonly experienced and familiar from an early age – i.e., the domain of animals, evidence from large-scale studies with students at many levels of education suggests that many students fail to acquire accurate knowledge about the composition of taxonomic categories (Trowbridge & Mintzes, 1985, 1988; Yen et al., 2004). The purpose of this study was to investigate the degree to which statistical regularities that are informative about the composition of taxonomic categories in the animal domain are present in source of input commonly experienced during development – i.e., children's books. Specifically, we investigated the degree to which animals from the same taxonomic category tended to co-occur on the same page of children's books.

Our findings indicate that current children's books about animals are not likely to provide a source of co-occurrence regularities that are informative about the composition of taxonomic categories. Specifically, books in which the number of same taxonomic category co-occurrences outnumber different taxonomic category co-occurrences could provide a source of input that fosters taxonomic category learning. Our analyses revealed that co-occurrences between members of the same taxonomic category primarily consisted of co-occurrences between members of the same species. Such same-species co-occurrences are likely to be less informative about the composition of taxonomic categories that co-occurrences between animals of different species from the same taxonomic category (e.g., Oakes et al., 1997; Perry et al., 2010). When co-occurrences between members of the same species are eliminated from the tally of same taxonomic category co-occurrences, the number of books in which same taxonomic category co-occurrences outnumber different taxonomic category co-occurrences is small.

These results suggest that children's books in the representative sample analyzed in this study do not provide a source of co-occurrence regularities that are informative about taxonomic category membership. Moreover, this deficit is particularly pronounced in books written for young children. The lack of exposure to informative co-occurrence regularities in early childhood may play a critical role in the emergence of persistent inaccurate conceptions of taxonomic category composition, given that recent research suggests early childhood is the period during which taxonomic misconceptions in the domain of animals are formed (Allen, 2015). These findings are consistent with the possibility that, even in a domain in which children receive a great deal of exposure from an early age, the input children receive from their environment may not provide them with co-occurrence regularities that can readily scaffold taxonomic category knowledge.

Open Questions & Future Directions

To date, the sources of environmental input of interest in research on taxonomic category acquisition have primarily consisted of visual similarity (Kloos & Sloutsky, 2008; Quinn, Eimas, & Rosenkrantz, 1993; Younger & Cohen, 1983) and shared labels (Fulkerson & Waxman, 2007; Robinson & Sloutsky, 2007). In contrast, the contributions of statistical regularities such as co-occurrence remain the subject of less direct study. Instead, multiple lines of evidence have strongly *suggested* a role for statistical regularities in shaping the way that children learn to group entities. For example, in the domain of animal knowledge, multiple studies have investigated the organization of children's knowledge about animals into groups by analyzing the order in which children spontaneously list animals when asked to say all the animals that come to mind. Findings from these studies have revealed that, across childhood, animals that appear in close sequential order in lists are those that co-occur in the same habitat (e.g., animals that co-occur in farms, water, zoos, etc.). Similarly, a series of studies conducted by Kattmann (Kattmann, 1998, 2001) revealed that, when prompted to organize animals into groups, children at multiple levels of education (ranging from ~9 to 13 years of age) made grouping decisions based on which animals typically co-occur. Finally, recent studies by Unger et al. (2016) have revealed that children perceive members of the same taxonomic category as related both earlier and to a greater extent when they commonly co-occur in the environment, versus when they do not.

Taken together, these findings support the possibility that a sensitivity to statistical co-occurrence regularities contributes to the way in which children's knowledge about the world becomes organized into groups or categories. However, a more direct investigation of this possible role for co-occurrence sensitivity ideally consist of research in which children are exposed to co-occurrence regularities, and the effects of this exposure on category knowledge are measured. For example, a follow-on to the present research might involve comparing the effects on children's knowledge of taxonomic animal categories of reading existing books with high versus low same taxonomic category co-occurrences, or a books developed in-lab to convey high versus low same taxonomic category co-occurrences while keeping other characteristics constant.

Conclusions

Sensitivities to statistical regularities in the environment may substantially contribute to our development of knowledge about the world around us. A key part of understanding the degree to which such contributions do indeed transpire is to assess the degree to which environmental input actually contains statistical regularities relevant to a given facet of knowledge about the world, such as the division of entities

into meaningful categories. Our present investigation yielded evidence that, even in a source of environmental input that is commonly experienced in childhood - i.e., children's books, and in a domain that is familiar from an early age – i.e., animals, statistical regularities that are informative about the composition of meaningful categories are relatively rare.

Acknowledgements

This work was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education, Institute of Education Sciences (R305B040063) and by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition – Scholar Award (220020401) to the second author. We also thank Anna Vande Velde, Eden Hu, Rebeka Almasi, Clara Lee, Camille Warner, Smriti Chauhan, and Sara Jahanian for their vital contributions to this research.

References

Allen, M. (2015). Preschool children's taxonomic knowledge of animal species. *Journal of Research in Science Teaching, 52*, 107-134.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition, 121*, 127-132.

Crowe, S. J., & Prescott, T. J. (2003). Continuity and change in the development of category structure: Insights from the semantic fluency task. *International Journal of Behavioral Development, 27*, 467-479.

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6-and 12-month-olds. *Cognition, 105*, 218-228.

Ganea, P. A., Ma, L., & DeLoache, J. S. (2011). Young children's learning and transfer of biological information from picture books to real animals. *Child Development, 82*, 1421-1433.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7*, 569-592.

Huebner, P., & Willits, J. (2017). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech.

Kattmann, U. (1998). Do students have an implicit theory of animal kinship. In B. Anderson (Ed.), *Research in Didaktik of Biology* (pp. 61-83). Göteborg, , Sweden.

Kattmann, U. (2001). Aquatics, Flyers, Creepers and Terrestrials—students' conceptions of animal classification. *Journal of Biological Education, 35*, 141-147.

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General, 137*, 52.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91-121.

Oakes, L. M., Coppage, D. J., & Dingel, A. (1997). By land or by sea: the role of perceptual similarity in infants' categorization of animals. *Developmental Psychology, 33*, 396.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical Learning in a Natural Language by 8-Month-Old Infants. *Child development, 80*, 674-685.

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally exemplar variability supports higher-order generalization and word learning. *Psychological Science, 21*, 1894-1902.

Pinkham, A. M., Kaefer, T., & Neuman, S. B. (2014). Taxonomies Support Preschoolers' Knowledge Acquisition from Storybooks. *Child Development Research, 2014*.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception, 22*, 463-475.

Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy, 11*, 233-253.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology, 22*, 1622-1627.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience, 16*, 486-492.

Storm, C. (1980). The semantic structure of animal terms: A developmental study. *International Journal of Behavioral Development, 3*, 381-407.

Trowbridge, J. E., & Mintzes, J. J. (1985). Students' alternative conceptions of animals and animal classification. *School Science and Mathematics, 85*, 304-316.

Trowbridge, J. E., & Mintzes, J. J. (1988). Alternative conceptions in animal classification: A cross-age study. *Journal of Research in Science Teaching, 25*, 547-571.

Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., & MacLellan, C. J. (2016). Developmental Changes in Semantic Knowledge Organization. *Journal of Experimental Child Psychology, 146*, 202-222.

Yen, C.-F., Yao, T.-W., & Chiu, Y.-C. (2004). Alternative conceptions in animal classification focusing on amphibians and reptiles: A cross-age study. *International Journal of Science and Mathematics Education, 2*, 159-174.

Young, K. T., Davis, K., Schoen, C., & Parker, S. (1998). Listening to parents: a national survey of parents with young children. *Archives of Pediatrics & Adolescent Medicine, 152*, 255-262.

Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development, 54*, 858-867.