Reasoning from Samples to Populations: Children Use Variability Information to Predict Novel Outcomes.

Elizabeth Lapidow (elapidow@ucsd.edu)
Department of Psychology, University of California, San Diego

Mariel Goddu (marielgoddu@fas.harvard.edu)
Department of Psychology, Harvard University

Caren M. Walker (carenwalker@ucsd.edu)
Department of Psychology, University of California, San Diego

Abstract

The ability to infer general characteristics of populations from specific instances is critical for reasoning. While there is evidence of this capacity in infancy, prior work has not examined children's ability to *use* these second-order inferences to make predictions about future outcomes. In the current study, 3-year-olds observed balls drawn at random from two containers. In one sample each ball was a different color. The other sample consisted of balls of only one (Experiment 1) or two (Experiment 2) colors. Children were asked which of the containers was more likely to contain a novel colored ball. A significant majority of children chose the more variable sample's container. This suggests that 3-year-olds are not only able to make inferences about hidden populations from the variability of observed samples, but also *use* those inferences to reason beyond their direct experience.

Keywords: cognitive development; variability; over-hypotheses; inductive inference

Introduction

A critical feature of human cognition is our ability to reason beyond the limits of our direct experience. From a small number of specific instances, we can make general inferences about the characteristics of populations, which in turn guide our future predictions and actions.

To illustrate, imagine you are looking into the window displays of several shops on a main street. In one, you see hats, shoes, and gloves. In another, cakes, breads, and jams. In a third, clocks, shovels, and paintbrushes. From these observations, you could readily infer that the first shop carries clothes, the next carries food, and the third carries tools – and that shops, in general, carry items that are similar in kind. This kind of inference, which draws conclusions at multiple levels of abstraction from the same information, is what Goodman (1955) terms an *overhypothesis*. In contrast with first-order inferences about the concrete properties of objects (e.g., recognizing the individual items in a window as '*hat,*' '*shoe,*' '*glove,*' and so on), overhypotheses are based on second-order properties, which capture the abstract relations between objects (e.g., recognizing that all the items belong to the same higher-order category, '*clothing*').

Utilizing Second-Order Inferences in Reasoning

This ability to form abstract knowledge from limited data is critical for human learning and reasoning. Overhypotheses impose constraints on subsequent inferences, allowing learners to make robust generalizations from relatively few observations (for examples from word learning, see Smith et al., 2002; Xu et al., 2012). Such inferences also facilitate reasonable predictions about events that the learner has never directly observed (Kemp et al., 2007). For instance, imagine that you've come to the shops in our example because you hope to purchase an umbrella. Only considering first order information about the items in each shop window (e.g., 'shoes,' 'cakes,' clocks') offers no guidance, since there are no umbrellas on display. However, if you are able to recognize the second-order relations among the objects that *are* on display (e.g., 'clothing,' 'food,' 'tools') and make inferences about the concealed populations from which they were drawn (the actual selection of merchandise offered), this supports the prediction that the 'clothing' and 'tools' shops are far more likely to sell umbrellas than the 'food' shop.

The current study asks whether this ability to use inferences about unobserved populations to form predictions about the likelihood of unobserved events is present in early childhood. There is some evidence of the capactiy to *form* overhypotheses in infancy. For example, Dewar and Xu (2010) showed 9-month-olds events in which four objects were drawn, apparently at random, from four identical boxes. The objects drawn from each of the first three boxes were all identical in shape (e.g., triangles drawn from the first box, squares from the second box, and circles from the third box). The experimenter then drew a single object (e.g., a star) from the fourth box, after which, one of two things happened: Either the next object drawn was the same shape (e.g., another star) or a different shape (e.g., a rectangle). Infants who saw the non-matching shape looked longer than those who saw the matching shape. This suggests that they had formed an overhypothesis—namely, that boxes contain objects of the same shape—and were therefore surprised when a non-matching shape was drawn, violating this expectation.

However, because this prior work employed *reactive* rather than *predictive* measures, it remains unknown whether young learners are able to actively *use* these second-order inferences to guide their subsequent reasoning. Returning to our shopping example: Simply forming general expectations about the shops from specific observations is not enough to determine where to search for an umbrella. That conclusion can only be reached by reasoning *prospectively* from those second-order inferences; employing them as the basis for a further inference about which shop is most likely to contain the unobserved item.

The current study asks whether 3-year-olds are capable of appropriately applying second-order inferences to inform their subsequent judgments. To do so, we showed children four balls, sampled at random, from each of two identical opaque containers. From one container, the experimenter drew a *four-color sample,* which consisted of four differently colored balls. From the other container, the experimenter drew either a *one-color* (all four balls were the same color, Experiment 1) or *two-color sample* (all but one of the four balls were the same color, Experiment 2). Children were then asked to make a prediction about the unobserved populations from which these samples were drawn. Specifically, children were asked which of the two containers they believed had a novel colored ball inside. This is a critical departure from the previous looking-time research, which inferred children's expectations from their *reactions* to incongruous observations (e.g., Dewar & Xu, 2010; Xu & Garcia, 2008). The current task not only assesses children's inferences, but also tests whether they can actively *use* them to generate explicit predictions about previously unobserved events.

Second-Order Inferences about Variability

Whether children make the appropriate prediction in our task will depend on their ability to infer the *variability* of the unobserved populations. The capacity to reason about properties like 'variability' is important for forming more abstract generalizations. That is, unlike overhypotheses that draw on higher-order *conceptual* knowledge (e.g., inferring that all of the items in the first shop window are 'clothing'), second-order properties like '*variable'* and '*uniform'* are entirely abstract, domain-independent features of events. Representing our observations in this way imposes powerful and potentially critical constraints on learning. Specifically, identifying the abstract characteristics of a problem allows learners to narrow their search to the subset of hypotheses in line with those features (see, Chu & Schulz, 2020; Schulz, 2012). For example, by preschool, children are able to evaluate candidate causes by matching their distributional (e.g., relative proportion) or dynamic (e.g., discrete/continuous, monotonicity/periodicity) properties to those of the outcome they observe (Magid et al., 2015; Tsividis et al., 2015). When do young learners begin to utilize these abstract features to guide their inferences, predictions, and actions?

There is some evidence that children recognize variability in both samples and populations. For instance, Denison and colleagues (2006) found that 4-year-olds are able to identify which of two populations is more likely to produce a random sample from their observation of the relative proportions of different objects in each. In a study of much younger children, Sim and Xu (2013) showed 8-month-olds four balls sampled (with replacement) from a box that was previously observed to contain six different colored balls. Infants looked longer and were more likely to explore the box when the sample drawn was uniform (i.e., four yellow balls), than when it was varied (i.e., a red, a green, a blue, and a yellow ball). In both studies, learners' behavior suggests they are sensitive to the variability of the populations and (correctly) expected this characteristic to be preserved in random samples. However, since participants always had an opportunity to observe the populations directly, these studies do not indicate whether young learners would be able to infer the variability of a population from samples alone.

In the current task, children never see the objects inside the containers, and they must therefore infer the variability of those unseen populations to inform their subsequent inference. If children only consider their observations in terms of first-order properties, then we would not expect them to show a preference for one container over the other. Neither sample includes a ball like the one that they are asked to make a prediction about, so neither population is more or less likely to contain it based on this information alone. On the other hand, considering the second-order properties of the samples readily leads to an inference about the unseen populations involved. Just as inferring the higher-order categories of the items displayed enables the window-shopper to make a prediction about where to buy an unseen item, inferring the relative *variability,* or *heterogeneity,* of the two samples should enable the learner to make a prediction that the *four-color sample* container is more likely to hold a novel-colored ball. If children preferentially select the *four-color sample* container, this would demonstrate that they have not only formed this second-order inference, but are also able to use it to guide their predictions and actions beyond the limits of their direct experience.

Experiment 1

The experimental design and analysis for this study was preregistered prior to beginning data collection (see: https://aspredicted.org/blind.php?x=rb4jn6).

Method

Participants A total of 40 children (M = 40.12 months, SD = 5.12 months, range = 25.35 – 47.8 months) were included. Participants were recruited and tested individually at local museums in a primarily urban area.

A priori power analysis was performed to calculate the target sample size. Our effect size (h = 0.72) was based on results from Erb, Buchanan, and Sobel (2013), which

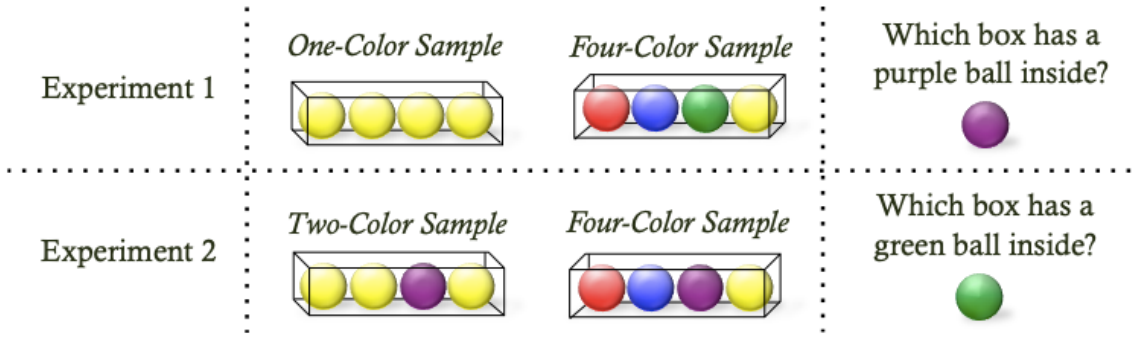

Figure 1: The task stimuli presented to participants in Experiment 1 (top row) and Experiment 2 (bottom row).

conducted a similar type of investigation (i.e., binomial analysis of a forced-choice question in which children were asked to match more or less complex mechanical insides to more or less variable machine functions) with a similar age group. This analysis indicated a minimum sample size of 38 to achieve a power of 0.8 at a significance level of 0.05. We rounded this minimum sample to 40 to accommodate counterbalancing.

Thirteen additional children were excluded and replaced due to experimental error (n = 2), sibling or caretaker interference (n = 6), or refusal to respond to the test question (n = 5).

Stimuli Two identical opaque containers were constructed from cardboard and painted black. Each container was 17" x 6" x 6" with a cardboard egg tray glued inside. This tray allowed the experimenter to arrange the balls inside each box in a specific order. The experimenter could then identify and select the correct order of balls without looking inside the box to give the illusion of random sampling. A felt-covered opening at the top of each box allowed the experimenter to reach inside and draw out the balls one at a time.

A total of ten plastic golf balls of different colors were used. These balls were placed inside of each of the two containers prior to the start of the task. One container held the *four-color sample*: four differently colored balls, one each of green, red, blue, and yellow. The other container held the *one-color sample* of four yellow balls. In addition, both containers also held one *novel ball*, which was purple.

The task also employed two 3" x 3" x 8" transparent plastic trays, which were used to hold the balls after they were drawn, and a photo of a single purple ball.

Procedure Each child participated one time. In person testing sessions began with the two opaque containers and clear trays on either side of the table. The experimenter told children they were going to play a game with the boxes, both of which had balls inside. She shook both containers so that the sound of the balls rattling inside was audible. After replacing the containers on the table, the experimenter said, "I am going to show you some of the balls in each box," and then stepped to one side so she was standing behind one of the two containers. The experimenter closed her eyes and turned her head away from the container while reaching in and pulling out a ball, apparently at random. She then directed her gaze towards the child while holding the ball out and saying, "Look!" After a beat, she placed the ball into the clear plastic tray beside the container. This process of "sampling" balls was repeated three more times, for a total of four balls in each sample. Afterwards, the experimenter stepped over to the other container and again drew four balls in the same manner.

In this way, each participant observed two sets of four balls drawn from the containers (see Figure 1, top row). The *one-color sample* consisted of four yellow balls, while the *four-color sample* consisted of one red, one green, one blue, and one yellow ball.[1] The balls in the *four-color sample* were always drawn from the box in the same order. The order and side of presentation of the samples were counterbalanced across participants.

After the second sample was drawn, the experimenter returned to the center of the table and addressed the child. Pointing at both the containers simultaneously, she said, "One of these two boxes has a purple ball, like this (holding up a card with a photograph of a purple ball), inside. Can you point to the box you think has the purple ball inside?" The card was then replaced face down on the table. If a child did not spontaneously indicate one of the two containers, the experimenter held up the picture card and pointed to each box, saying, "Do you think there is a purple ball in this one, or this one?" Children who did not respond after two such prompts were excluded and replaced.

We recorded whether each child chose to search for the novel-colored ball in the container that produced the *four-color* sample or the *one-color* sample. After children indicated their choice, the experimenter reached into the selected container and drew out a purple ball.

Results

A significant majority of children (72.5%) chose the *four-color sample* container (p = 0.006, two-tailed binomial). A

[1] Samples were selected based on the procedure used by Sim and Xu (2013).

logistic regression, treating age as a continuous factor and choice of the *four-color sample* container as the dependent variable, revealed no effects of age on final choice (Wald, $z = 0.881$, $p > 0.378$, *ns*).

These results suggests that young learners are not only able to form second-order inferences from the data they observe, but also make use of these inferences to reason beyond their direct experience. However, the experimental design used in Experiment 1 leaves open the possibility that this success was due to children's *avoidance* of the uniform *one-color* sample, rather than an inference about the relative variability of the two populations. In order to test this possibility in Experiment 2, we first designed and validated an online version of the procedure (Experiment 1a).

Experiment 1a

Data collection on Experiment 1 concluded just prior to lock-down restrictions resulting from COVID-19. Therefore, in order to conduct Experiment 2, it was first necessary to develop a version of this task that could be administered online, and validate it by replicating the results of Experiment 1. The methods and findings of this interim investigation are reported below (see https://aspredicted.org/blind.php?x=t85n33).

Method

Participants A total of 43 participants ($M = 41.74$ months, $SD = 3.47$ months, range = 36.2 – 47.9 months) were recruited via email from a database of families from the same population as Experiment 1, and tested via a Zoom video call.

Six additional children were excluded and replaced due to caretaker interference ($n = 1$), refusal to respond to the test question ($n = 1$), or technical issues such as interruption by an unstable internet connection ($n = 4$).

Stimuli and Procedure Participants watched a series of prerecorded videos of an experimenter performing the procedure from Experiment 1. In order to ensure that the ball colors would be clearly distinguishable across different computer monitors: a purple ball replaced the green ball in the *four-color sample* and the novel-ball was green instead of purple.

Experiment 1a was presented in a series of prerecorded videos via Slides.com. These videos are available via OSF: https://osf.io/5x8ku/?view_only=269c5468936d4811a55f237041f9ff96. The only difference from the procedure in Experiment 1 was the addition of a pre-trial task, in which children were asked to point to a black triangle that appeared first in the left and then the right side of the screen (order counterbalanced). This gave participants a chance to practice responding by pointing in the online task and provided a visual calibration for the experimenter to determine which container was chosen at test.

Results

Children's online performance showed a similar, but weaker pattern as the one observed in Experiment 1: only 62.79% of children in Experiment 1a selected the *four-color sample* container. Although this proportion was not significantly different from Experiment 1 ($p = 0.213$, two-tailed binomial), it was also not significantly different from chance ($p = 0.126$, two-tailed binomial).

A post-hoc analysis assessed whether this partial-replication might be due to age-related differences associated with the change in modality. There was no significant difference in age between the participants tested in Experiments 1 and 1a, $t(81) = -1.7$, $p = 0.09$ (*ns*) and a logistic regression treating age as a continuous factor was again not significant (Wald, $z = 1.271$, $p > 0.204$, *ns*). However, a median-split revealed a clear age difference in online performance. Younger children ($n = 21$, $M = 38.62$ months, $SD = 1.74$ months, range = 36.2 – 41.5 months) selected the *four-color sample* container 52.38% of the time, which was significantly less often than older children ($n = 22$, $M = 44.71$ months, $SD = 1.49$ months, range = 42.21 – 47.9 months), who selected this container 72.73% of the time ($p = 0.048$, two-tailed binomial). Importantly, this age effect was *not* found in children tested in person: younger children in Experiment 1 ($n = 18$, $M = 35.61$ months, $SD = 4.13$ months, range = 25.3 – 40 months) selected the *four-color sample* container 66.66% of the time, which was not significantly different from older children ($n = 22$, $M = 43.8$ months, $SD = 1.81$ months, range = 41.4 – 47.8 months), who selected this container 77.27% of the time ($p = 0.269$, two-tailed binomial). This suggests that the difference observed between Experiments 1 and 1a is likely due to the poorer performance of younger children in the online modality. For further discussion of the interaction between modality and age, see Lapidow et al. (2021).

Discussion

The results of Experiment 1 and 1a provide initial evidence that young learners form abstract hypotheses about populations from the characteristics of the samples they observe, and apply them to guide their subsequent inferences and actions. However, as noted above, one alternative interpretation of these results is that children succeeded by *avoiding* the container that produced the uniform *one-color sample,* rather than by *selecting* the container that produced the variable *four-color sample*.

A low-level, perceptual version of this alternative can be largely dismissed. That is, children could not simply compare the first-order properties of each sample (i.e., the colors of the balls) and reject the sample of yellow balls because it does not match the color of the novel ball. This cannot account for children's choice behavior, since, considered only in terms of these first-order properties, the two samples are equivalent in failing to match the color of the novel ball. Even representing the *one-color sample* as "all yellow" requires forming a second-order inference about the uniformity of the colors observed. As such, we

can be confident that children's predictions reflect the formation of an overhypothesis about a second-order property.

Nevertheless, because the samples used in Experiment 1 suggest populations that are completely uniform (*one-color*) or varied (*four-color*), these results alone are insufficient to demonstrate an inference about variability in particular. It is possible to arrive at the correct prediction in this task by inferring that the *one-color sample* was drawn from a *uniform* population (and, therefore, rejecting it), without reference to the other. Uniformity and variability are, of course, closely related concepts, and both are examples of the kind of second-order abstract properties that could provide critical constraints on hypothesis search. However, determining which of these approaches underlies young children's ability to draw predictions about novel outcomes from their second-order inferences requires presenting participants with less extreme samples. Specifically, we designed a follow-up experiment in which the *one-color sample* of four yellow balls was replaced with a *two-color sample* of three yellow balls and one non-yellow ball.

This is a considerably more challenging version of the inference problem from Experiments 1 and 1a. For one, there is greater first-order perceptual similarity between the samples (which now have two colors in common). Moreover, the difference between the populations that can be inferred from these samples is also subtler: children must distinguish between *more variable* and *less variable*, rather than between *variable* and *uniform*. Furthermore, since neither of the options suggests uniformity, children cannot succeed by simply avoiding it. Instead, they have to compare the relative variability of by each sample in order to arrive at a prediction about the relative variability of the populations.

Experiment 2Experiment 2 aimed to test whether children's success on Experiment 1 was due to their avoidance of the uniform container. Children observed a *four-color sample* and a *two-color sample* drawn from two concealed populations. They were then asked to make a prediction about which of the two populations contains a novel-color ball. If children again select the *four-color sample*, this would suggest that their predictions were informed by a second-order inference about the relative variability of each population.

Given the interaction between age and online modality observed in Experiment 1a, only older 3-year-olds (3.5 to 4.0) were included in Experiment 2.

Method

Participants Twenty participants (M = 42.92 months, SD = 1.26 months, range = 42.2 – 47.7 months) were recruited in the same manner as Experiment 1a.

Six additional children were excluded due to caretaker interference (n = 1), experimenter error (n = 2), or because of technical issues interrupting the session (n = 3).

Stimuli & Procedure The materials and procedure was identical to that used in Experiment 1a with one exception: the samples included a *four-color sample* consisting of a red, a blue, a purple, and a yellow ball (drawn in that order) and a *two-color sample* consisting of three yellow and one purple ball (drawn as yellow, purple, yellow, yellow). As in Experiment 1a, the *novel ball* was green (see Figure 1, bottom row).

Results

A significant majority of children (80%) chose the *four-color sample* container (p = 0.01, two-tailed binomial). A logistic regression, treating age as a continuous factor and choice of the *four-color sample* container as the dependent variable, revealed no effects of age on final choice (Wald, z = 0.448, p > 0.654, *ns*). These results demonstrate that children's responses are not due to their avoidance of uniformity, but their prediction about the relative variability of the two unseen populations

General Discussion

Together, these experiments provide evidence of a critical but often overlooked aspect of the early development of higher-order cognition: the ability to infer and apply second-order inferences to guide future reasoning. We queried young children's ability to reason beyond their immediate observations by asking for a judgment about which of two unseen populations was more likely to contain an item they had never seen before. Across two experiments, a significant majority of children selected the *four-color* sample over the *one-color* sample (Experiment 1) and *two-color* sample (Experiment 2). If children's predictions were based only on first-order information, or if children were unable to accurately apply their second-order inferences, then we would not expect them to have this strong intuition. After all, there is no objectively correct answer to this question; children have no factual evidence to rule out the possibility that the novel object is inside the less variable of the two containers. Instead, the results suggest that children actively compare their second-order inferences about each sample to form a conclusion about the relative variability, or *heterogeneity*, of the unseen populations. This relative estimate is, in turn, used to predict which population was more likely to contain a novel item.

Our findings compliment and extend the existing developmental research on the emergence of second-order inferences. Past studies suggest that even infants would have been surprised if the novel-colored ball had been drawn from the less variable of the two containers (e.g., Dewar & Xu, 2010; Xu & Garcia, 2008). We also know that infants can reason about the relationship between populations and samples, but only for situations where the population has been directly observed (e.g., Sim & Xu, 2013). The current studies go beyond this prior work to show that children are also able to make inferences about unseen populations from samples, and use this to make a prediction about an outcome that has not yet occurred. To

our knowledge, only one other study has examined children's use of overhypotheses about concealed populations, and reports only mixed success. Specifically, Felsche and colleagues (2019) found that 4- and 5-year-olds' choices to select prizes from different containers were sensitive to the characteristics of previously observed samples, but only when prize type (and not size) was the critical cue. This suggests that children may have only been attending to this single dimension, rather than fully representing the second-order population characteristics implied by the samples. In contrast, 3-year-olds' successful prediction of the likely location of a novel object in the current task provides clear evidence for their ability to utilize abstract inferences to guide subsequent actions.

Our findings also offer an interesting connection to research on the development of abstract relational reasoning more generally. While research indicates that toddlers (18-30-month-olds) infer and apply second-order *same-different* inferences to inform their subsequent causal judgments (e.g., Goddu et al., 2021; Walker & Gopnik, 2013), 3-year-olds typically fail (Walker et al., 2016). Instead, these older children tend to privilege first-order properties, often preventing them from recognizing second-order relations, even when given explicit instructions to do so (Christie & Gentner, 2010). It is intriguing, therefore, that 3-year-olds in the current study did not reason about the samples in terms of their first-order properties, and instead successfully generated and utilized abstract inferences. More research is needed to better understand the connection between children's inferences about relative variability, and their ability to detect same-different in higher-order relational reasoning tasks.

Future work may also aim to elaborate and expand upon the current findings by investigating children's performance when the task is presented in contextualized, real-world domains. This question is intriguing since, despite evidence that learners form appropriate inferences from variability information as early as 8- or 10-months (e.g., Dewar & Xu, 2010; Sim & Xu, 2013; Xu & Garcia, 2008), young children often fail when provided with contextualized versions of similar questions. Erb, Buchanan, and Sobel (2013), for example, find that although 4-year-olds recognize variability information when reasoning about the diversity of functions performed by a machine and use this information as a basis for subsequent inferences, 3-year-olds do not. In fact, in a related study, Ahl and Keil (2017) asked children to match the diversity and number of machine functions to the complexity of their insides, and found no evidence of such inferences before the age of six. It is possible that equating the variability of functions with the complexity of mechanisms is more difficult than making inferences about future novel outcomes, regardless of the surface-level content. But it is also possible that the ease with which young learners recognize and reason from second-order properties like variability differs across domains.

Our direct experience of the world is limited—but our ability to reason about these data at multiple levels of abstraction allows us to go beyond these observations. Here, we show that young children readily infer second-order properties of unseen populations from samples, and accurately utilize those inferences to guide their reasoning about novel outcomes. These initial findings help to support our understanding of how human learning allows us to reason beyond the limit of our direct experience and lays the foundation for future research exploring how this capacity influences everyday reasoning.

Acknowledgments

The authors would like to thank Tushita Tandon for her invaluable help in data collection and management of this project, Kim Scott and the Lookit Team for their tireless support of online research, the CRADLE team behind ChildrenHelpingScience.com, Trisha Katz, Phoebe Betts, Xiaoyang Chu, and Amberley Stein for their assistance with data collection, and the participating children and families, both in-person and online.

Funding for this project was provided by the Jacobs Foundation, an NSF Career Award (SBE #2047581) to C. M. Walker, and the National Defense Science and Engineering Graduate Fellowship to E. Lapidow.

References

Ahl, R. E., & Keil, F. C. (2017). Diverse Effects, Complex Causes: Children Use Information About Machines' Functional Diversity to Infer Internal Complexity. *Child Development*. https://doi.org/10.1111/cdev.12613

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*. https://doi.org/10.1080/15248371003700015

Chu, J., & Schulz, L. E. (2020). Play, Curiosity, and Cognition. *Annual Review of Developmental Psychology*. https://doi.org/10.1146/annurev-devpsych-070120-014806

Denison, S., Konopczynski, K., Garcia, V., & Xu, F. (2006). Probabilistic reasoning in preschoolers: random sampling and base rate. *Proceedings of the …*.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science*. https://doi.org/10.1177/0956797610388810

Erb, C. D., Buchanan, D. W., & Sobel, D. M. (2013). Children's developing understanding of the relation between variable causal efficacy and mechanistic complexity. *Cognition*. https://doi.org/10.1016/j.cognition.2013.08.002

Felsche, E., Stevens, P., Völter, C., Buchsbaum, D., & Seed, A. (2019). Exploring the use of overhypotheses by children and capuchin monkeys. *CogSci*, 1731–1737.

Goddu, M. K., Sullivan, J. N., & Walker, C. M. (2021). Toddlers learn and flexibly apply multiple possibilities. *Child Development*, *92*(6), 2244–2251.

Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard

University Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*. https://doi.org/10.1111/j.1467-7687.2007.00585.x

Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A Tale of Three Platforms: Investigating Preschoolers' Second-Order Inferences Using In-Person, Zoom, and Lookit Methodologies. *Frontiers in Psychology*, *12*.

Magid, R. W., Sheskin, M., & Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development*. https://doi.org/10.1016/j.cogdev.2014.12.008

Schulz, L. E. (2012). Finding New Facts; Thinking New Thoughts. *Advances in Child Development and Behavior*. https://doi.org/10.1016/B978-0-12-397919-3.00010-1

Sim, Z., & Xu, F. (2013). Infants' Early Understanding of Coincidences. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*. https://doi.org/10.1111/1467-9280.00403

Tsividis, P., Tenenbaum, J. B., & Schulz, L. E. (2015). Hypothesis-Space Constraints in Causal Learning. *Annual Meeting of the Cognitive Science Society*.

Walker, C. M., Bridgers, S., & Gopnik, A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring abstract concepts. *Cognition*, *156*, 30–40.

Walker, C. M., & Gopnik, A. (2013). Toddlers Infer Higher-Order Relational Principles in Causal Learning. *Psychological Science*, *25*(1), 161–169. https://doi.org/10.1177/0956797613502983

Xu, F., Dewar, K., & Perfors, A. (2012). Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism. In *The Origins of Object Knowledge*. https://doi.org/10.1093/acprof:oso/9780199216895.003.0011

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0704450105