Exploring the use of overhypotheses by children and capuchin monkeys

Elisa Felsche[1] (ef68@st-andrews.ac.uk),
Patience Stevens[2] (pstevens@andrew.cmu.edu),
Christoph Völter[3] (christoph.voelter@vetmeduni.ac.at),
Daphna Buchsbaum[4*] (buchsbaum@psych.utoronto.ca)
and Amanda Seed[1*] (ams18@st-andrews.ac.uk)

[1]School of Psychology and Neuroscience, University of St Andrews, Scotland
[2]Department of Psychology, Carnegie Mellon University, USA
[3]Messerli Research Institute, University of Veterinary Medicine Vienna, Austria
[4]Department of Psychology, University of Toronto, Canada

Abstract

The use of abstract higher-level knowledge (overhypotheses) allows humans to learn quickly from sparse data, and make predictions in new situations. Previous research has suggested that humans may be the only species capable of abstract knowledge formation, but this remains controversial, and there is also mixed evidence for when this ability emerges over human development. Kemp et al. (2007) proposed a computational model of overhypothesis formation from sparse data. We provide the first direct test of this model: an ecologically valid paradigm for testing two species, capuchin monkeys (Sapajus spp.) and 4-5-year-old human children. We compared performance to predictions made by models with and without the capacity to learn overhypotheses. Children's choices were consistent with the overhypothesis model predictions, whereas monkeys performed at chance level.

Keywords: Overhypotheses, abstraction, generalization, animal cognition, computational modeling, cognitive development

Introduction

For long-lived species that exploit a complex environment it might be beneficial to transfer adaptive behavior across situations, through the formation of abstract generalizations. For example, if a primate learns that one tree grows figs, a second papaya and a third nuts, at a more abstract level she is also exposed to the regularity: "Trees carry a uniform fruit type". Learning this abstraction would make just one bite of fruit from a new tree sufficient to decide whether or not continued foraging in this tree would be beneficial.

In the developmental literature the term 'overhypotheses' (Goodman, 1955) describes such higher-order generalizations at an abstract level that inform inferences about more specific hypotheses (Kemp, Perfors, & Tenenbaum, 2007). Kemp et al. (2007) developed a computational model that suggested that, in principle, overhypotheses can be learned quickly from sparse data and used to make wide-ranging predictions in new situations.

Evidence for a possible early emergence of this ability during human infancy comes from a study using looking-time methodology. Dewar and Xu (2010) presented 9-month-olds with sampled evidence supporting the overhypothesis that containers are filled with objects of the same shape. In a test situation, infants looked longer when two differently shaped objects were drawn from the same container, contradicting this overhypothesis, than when two uniformly shaped objects were sampled.

Despite this evidence for early overhypothesis formation, other methods show contrasting results. A common method to assess understanding of the abstract concepts "same" and "different" is the relational matching-to-sample (RMTS) task. Here, participants are presented with an example stimulus pair (either two of the same or two different items) as well as two test pairs, and must select the pair with the matching abstract relation to the example. Hochmann et al. (2017) showed that children begin to succeed in a 2-item RMTS task by the age of 5 but not earlier (see Kotovsky & Gentner, 1996 for a similar result). However, labeling the relations verbally enables children to succeed in the RMTS task as early as age 2 (Christie & Gentner, 2014).

In contrast, in an anticipatory looking time procedure, Hochmann, Carey and Mehler (2018) showed that 7 and 12-month-olds were sensitive to the abstract relation of same but not different. Similarly, 18- to 30-month-old children correctly selected either a matching or a dissimilar pair of objects following evidence that their relation was causally relevant (Walker & Gopnik, 2014).

In addition, only a few non-human species master the RMTS task, usually after lengthy training regimes (e.g. Truppa, Mortari, Garofoli, Privitera, & Visalberghi, 2011; see also Smirnova, Zorina, Obozova & Wasserman (2015)), and often only with multi-stimulus arrays instead of stimulus pairs (see Wasserman & Young, 2010 for a review; the latter also helping 3-year-olds succeed, Hochmann et al., 2017). As a result, some have suggested that the RMTS task can be solved by perceptual processes alone, and that abstract knowledge is a uniquely human capability (Penn, Holyoak, & Povinelli, 2008; Vonk, 2015). In a different set of tasks, chimpanzees and bonobos have been suggested to use relative spatial relations such as "top" or "middle" to find hidden food rewards (Haun & Call, 2009; Christie, Gentner, Call & Haun, 2016). However, it is not clear whether searching based on relative rather than absolute

[*] Equal contribution

spatial relations represents the same kind of abstract knowledge as concepts such as "same" and "different". In summary, the question of whether abstract knowledge formation is an evolutionary primitive, shared with other species and emerging early in human development, or a recently-evolved, late-developing skill, is complicated by considerable methodological differences between the tasks used across ages and species. Further, as in other areas of cognitive development, there is something of a dissociation between looking time results that suggest an early-emerging conceptual competence, and later emerging success on choice-based measures by older children. One concern is that successful discrimination in infant looking time tasks may not require the same kind of conceptual competence as paradigms requiring participants to use their knowledge to make a choice (e.g., Hood, 2004).

We therefore designed a task that could be used across species, to examine abstract knowledge formation in an ethologically valid context without extensive training or explanation, based on the original idea of overhypothesis formation by Goodman (1955). Importantly this allowed us to test a theoretical computational model for how limited data can be sufficient for overhypothesis formation in this task (Kemp et al., 2007). Similar to Dewar and Xu's (2010) infant looking time study, we adapted Goodman's thought experiment, in which bags of marbles can be either uniform or mixed in color, to create a choice paradigm suitable for older children and capuchin monkeys. We presented sampled evidence from three containers either supporting the overhypothesis that rewards are sorted by their *size* or by their *type*. At test, participants were presented with two new containers and one example item from each: a small, high-valued reward from A and a large, low-valued reward from B (Figure 1). Participants then chose between two covert samples from these new containers. Differential choice between conditions—namely, choosing A to obtain a high-valued option in the type condition, but choosing B to obtain a large item in the size condition—would reflect sensitivity to the overhypotheses governing object sorting.

Computational Model

Probabilistic hierarchical Bayesian models have frequently been proposed as computational models of children's rapid early learning (Kemp et al., 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). They demonstrate how, in principle, knowledge can be acquired at multiple levels of abstraction simultaneously, after only seeing small amounts of data. Kemp et al. (2007) show how more abstract hypotheses can constrain the hypothesis space at lower levels, leading to rapid inferences when encountering new but related situations. Due to the interdependence of concrete observations and higher-order concepts, these models do not exhibit the tension between low-level and higher-level learning often discussed in the animal literature. However, while the Kemp et al. model has successfully characterized existing findings in the developmental literature, the model's predictions have not been directly empirically tested in children or animals.

Here, we extended the Kemp et al. (2007) model with a rational choice rule, allowing us to directly compare the model's predictions for which test container (A or B) learners should choose to receive a reward from with new empirical data. We infer the relative utilities of the different reward types, based on the participants' choices in preference testing, following the inverse preference model developed by Lucas et al. (2014).

Model Overview.

Figure 1 provides an overview of both our task and of the computational model. In this model, items are sampled from evidence containers, each of which has a distribution of items with different features (i.e., item type and size). These distributions capture a first level of abstract knowledge (level 1), describing the kinds of items likely to be found in this specific container. Simultaneously, the model also represents a more abstract level of knowledge (level 2), which describes the probability distribution over containers—the extent to which containers in general tend to be mixed or uniform, and the distribution of features across containers. Using this hierarchical structure, the model captures how specific observations of samples from individual containers can be used to simultaneously infer parameters at multiple levels of abstraction.

Learning Overhypotheses.

As in Kemp et al., (2007) we use a Dirichlet-multinomial model (Gelman, Carlin, Stern & Rubin, 2003). The individual sees evidence items $\boldsymbol{y^i}$ with d feature dimensions (in our case d = 2: the item's type and size), sampled from each evidence container i. We assume that items are drawn randomly and independently from each container and that the item's type is determined independently of its size. The item types (sizes) are sampled from $\boldsymbol{y_d^i} \sim$ Multinomial ($\boldsymbol{\theta_d^i}$), the distribution over item types (sizes) in that container. Each container's distribution over item types (sizes), $\boldsymbol{\theta_d^i}$, is in turn sampled from a Dirichlet distribution, parameterized by a scalar α_d and a vector $\boldsymbol{\beta_d}$, $\boldsymbol{\theta_d^i} \sim$ Dirichlet(α_d, $\boldsymbol{\beta_d}$). These hyperparameters characterize the overhypothesis across containers. α_d parameterizes the extent to which items in each container are uniform in type (size). $\boldsymbol{\beta_d}$ represents the type (size) distribution across the entire set of containers. α_d is in turn sampled from an exponential distribution, $\alpha_d \sim$ Exponential(1), and $\boldsymbol{\beta_d}$ from a symmetric Dirichlet distribution, $\boldsymbol{\beta_d} \sim$ Dirichlet(1).

To model overhypothesis formation, we infer $p(\alpha_d, \boldsymbol{\beta}_d | \boldsymbol{Y_d})$ (referred to as $p(\alpha, \boldsymbol{\beta} | \boldsymbol{Y})$ for simplicity below), the posterior distribution over (α, $\boldsymbol{\beta}$), given the observed items $\boldsymbol{y^i}$, drawn from the N evidence containers,

$$p(\alpha, \boldsymbol{\beta} | \boldsymbol{Y}) \propto \int \prod_{i=1}^{N} p(\boldsymbol{y^i} | \boldsymbol{\theta^i}) \boldsymbol{p}(\boldsymbol{\theta^i} | \alpha, \boldsymbol{\beta}) p(\alpha) p(\boldsymbol{\beta}) \, d\theta \qquad (1)$$

estimated using the Metropolis-Hastings algorithm. Here we used 5 chains with 2000 samples and a burn in of 1000.

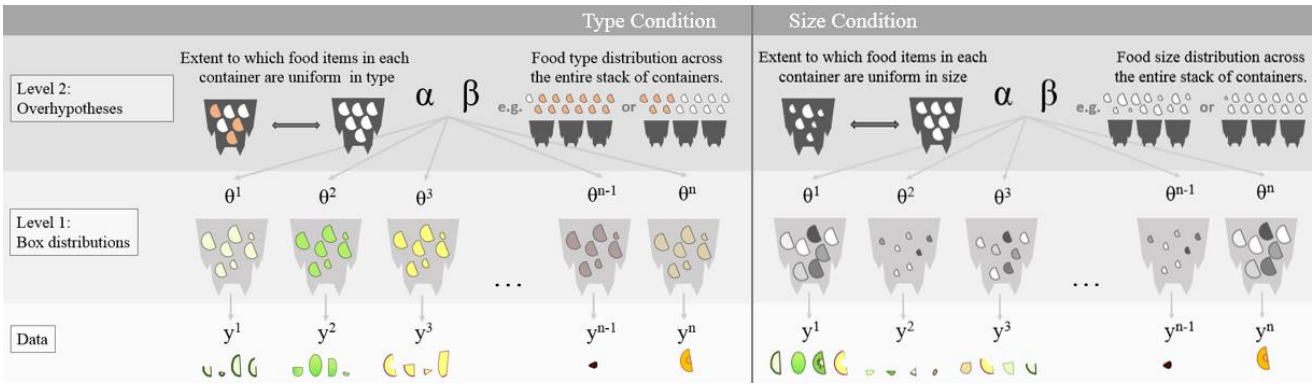

Figure 1: Hierarchical Bayesian model of overhypothesis formation. The parameters α and **β** describe an overhypothesis at the second level of abstraction: α represents the extent to which containers in general tend to be uniform for a given feature dimension, and **β** captures the feature variability across all containers. Feature distributions of a specific container ($\boldsymbol{\theta}^{\mathbf{i}}$, Level 1 abstraction), are constrained by overhypotheses at Level 2, and in turn constrain the items $\boldsymbol{y}^{\boldsymbol{i}}$ sampled from that container.

Predicting the content of the test buckets

We would like to predict the type (size) of the j^{th} (unseen) sample from the new test container $i+1$, given already known samples from this test container, -j (everything not j), and the overhypotheses inferred from the evidence containers. For a Dirichlet-Multinomial distribution, $p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}, \alpha, \boldsymbol{\beta})$, the posterior predictive distribution for the type (size) of the next item in the container, given the previously seen items and the hyperparameters $\alpha, \boldsymbol{\beta}$, has a simple closed form solution. Marginalizing over $p(\alpha, \boldsymbol{\beta}|\boldsymbol{Y})$, the posterior distribution over possible values of α and **β**, estimated from the evidence containers give us,

$$p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}) = \iint p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}, \alpha, \boldsymbol{\beta}) p(\alpha, \boldsymbol{\beta}|\boldsymbol{Y})\, d\alpha, d\boldsymbol{\beta} \qquad (2)$$

Approximated by averaging $p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}, \alpha, \boldsymbol{\beta})$ across sampled values of $p(\alpha, \boldsymbol{\beta}|\boldsymbol{Y})$.

Predicting choice of test item

Given the distribution over possible next items from each test container, we would like to predict the learner's choices. We assume that learners are choosing which box to take the next item from based on the expected utility of the next item from each container. As in Lucas et al. (2014), we assume that the utility of an item x is just the product of the utility of its individual features. For simplicity we assume that utility scales linearly with item size, s_x, so that the utility of item x, is $u_x = s_x \cdot \delta_{t_x}$, where δ_{t_x} is the learner's utility for one unit of item type t_x. The utility of a container is calculated by summing the utilities of each possible item, weighted by its probability of being the next item. As in previous work, we assume that learners become exponentially more likely to choose a container i as its expected utility increases.

$$P(c = i|u) = \frac{e^{u_i}}{\sum_j e^{u_j}} \qquad (3)$$

Inferring reward utilities.

To compute the relative utilities of the different reward items, prior to the main experiment, we conducted a series of preference tests, where participants chose which of two reward items they wanted. For simplicity, we only included the categorical item types high, medium and low-value. Comparisons included choices between different item types of fixed size, between different sizes of the same type, as well as mixed comparisons between large items of low value and small items of high value.

Following the preference inference model described in Lucas et al. (2014), we assume that learners choose items based on their relative utilities as in equation 3. We infer item type utilities **u** from learner's choices **c,** separately for each species, by computing the posterior probability $p(u|c) \propto p(c|u)p(\boldsymbol{u})$, estimated using the Metropolis-Hastings algorithm. Following Lucas et al. (2014), we assume that the type preferences **δ** are normally distributed, with $\mu = 0$, and variance $\sigma^2 = 2$ (however the inferred preferences are robust to different values of σ^2). Here we used one chain with 10000 samples and a burn in of 500.

Model Predictions.

Using this approach, we inferred strong preferences for high vs low value items for both species (children: Δ0.62; monkeys: Δ1.19). We used each species item utilities, separately inferred from their preference task data, to make *a priori* choice predictions for our experiment. Model predictions based on Level 2 abstraction (abstraction across containers) make clear contrasting choice predictions between the size and type conditions for both species after only one trial (one set of 3 evidences containers; Figure 2a). Predictions across subsequent trials, after seeing up to 6 sets of evidence containers get asymptotically more extreme (Figure 2b). In contrast, for a lesioned model capable of only Level 1 abstraction, and thus not learning from the evidence containers, the test container with the small, high value item is the preferred choice independent of condition.

Experiment 1: Abstraction across containers

Methods

Participants. Participants were 80 4- to 5- year-old children (M age = 4.9 yrs, 50% female), recruited at two local museums in Toronto, Canada. Eight additional children were excluded from analysis because they ended the game early (n = 5) or due to experimenter error (n=3). Seventeen brown capuchin monkeys (*Sapajus spp.*, M age = 6.5 yrs, 29% female) completed a preparatory food preference testing. Due to motivation decline only 11 monkeys finished the main study and are included in the data analysis.

Materials & Procedure. For the monkeys, nine different types of food items (divided into 3 categories: high, medium and low value) and 5 item sizes were used. Rewards for children were stickers picturing either animals (high value) or simple shapes (low value). Size was manipulated by the number of stickers on a strip, varying from 1 to 5. To encourage consistent sticker preferences across children, they were given the task of filling in a zoo map with as many animals as possible, making animal stickers more valuable than shape stickers. Prior to the main experiment, both species received preference testing, details of this procedure are given below. All sessions were video recorded.

Main Experiment. For both species the procedure in each trial was very similar. The experimenter successively sampled four example items from each of 3 evidence containers into transparent cups (monkeys), or onto metal frames (children), starting always on the left side. Depending on the condition, the items from one container were either all of the same type but of varying sizes (type condition) or all identical in size but different in type (size condition, see Figure 1). During the sampling, the experimenter closed her eyes and kept her head upright to create the illusion of random sampling.

Subsequently, two new test containers were brought forward, with the other containers and their evidence still in view of the participants off to the side. The experimenter first simultaneously sampled one evidence item from each test container. This was always a small, high-valued reward from container A and a large, low-valued reward from container B (item types counterbalanced). The experimenter then sampled another item from each container simultaneously, this time keeping the reward items hidden in her closed hands. The closed hands were extended towards the participants so that they could indicate their choice by reaching towards one of the hands. Participants were rewarded with the chosen item. Reward items were chosen to be in line with the condition overhypothesis (i.e., of the expected type or size), at least of medium value in the size condition, and otherwise randomly sampled.

For monkeys, the experimenter crossed her hands in half of the trials (a procedure they are familiar with) to ensure they tracked the hidden sample in the experimenter's hand and were not just pointing towards the sampled items. For children, hands were never crossed. In comparison to the monkeys, children's pointing was not restricted by a choice panel and thus they were able to clearly indicate a specific hand rather than only a side (unlike the monkeys children also had no prior experience with this procedure and showed confusion about the hands crossing in a pilot study).

Due to the small available sample, monkeys experienced both conditions, size and type, in a within-subject ABAB design, with the first condition counterbalanced across monkeys. Here, two different kinds of containers, bags and boxes (counterbalanced), were used, so that any overhypothesis could be tied to a specific kind of container. Monkeys received 16 sessions with 3 trials each, with 4 sessions per block. Children were tested in a between-subject design to allow us to test them in a single session in a science museum. and thus only presented with one container type (boxes). They received one session of 6 trials. Importantly, as for the monkeys, they did not receive any explicit instruction concerning the abstract rules governing the reward distribution.

Reward Preference Testing. Prior to the main experiment, we conducted preference testing to ensure that participants preferred bigger over smaller (size comparisons) and high over low-value rewards (type comparisons). Further small, high-value items were compared to large, low-value items (mixed comparisons). Monkeys received 9 kinds of size comparisons, one for each food type. There were also 6 kinds of type and mixed comparisons respectively, as each of the three high-value items was compared to two low-value items. Finally, the least liked high-value item was compared to all 3 medium valued items to ensure a clear preference. Monkeys received 10 trials for each of the 24 comparisons, presented over 24 sessions. Food items were presented in a covered forced choice procedure, where the monkeys first saw the food on the experimenter's palms and then had to choose between her closed fists.

Children first received a warm-up of 3 trials in which they were familiarized with the closed-hands choice procedure. Due to the constraints of museum testing, children were presented with a reduced preference procedure of two preference trials each for the type and size comparisons. A subset of n=58 children also received two mixed trials. Following preference testing, for the main experiment, novel stickers were used, and children were asked to find a lot of animals for a new, blank zoo map.

Results

Reward Preference Testing. In the type comparisons, both species significantly preferred high-value items over equally sized low-value alternatives (Capuchins: M = 0.86, SD = 0.12, $t(16)=11.78$, $p<0.001$; Children: (M = 0.94, SD = 0.18, $t(79) = 22.33$, $p < 0.001$). Capuchins further preferred the least liked high value item over equally sized pieces of medium-valued foods (M = 0.89, SD = 0.06, $t(16)=25.07$, $p<0.001$). Both groups also significantly preferred large over small items (Capuchins: M = 0.83, SD = 0.06, $t(16)=23.96$, $p<0.001$; Children: M = 0.83, SD = 0.32, $t(79) = 9.11$, $p < 0.001$). In the mixed comparisons

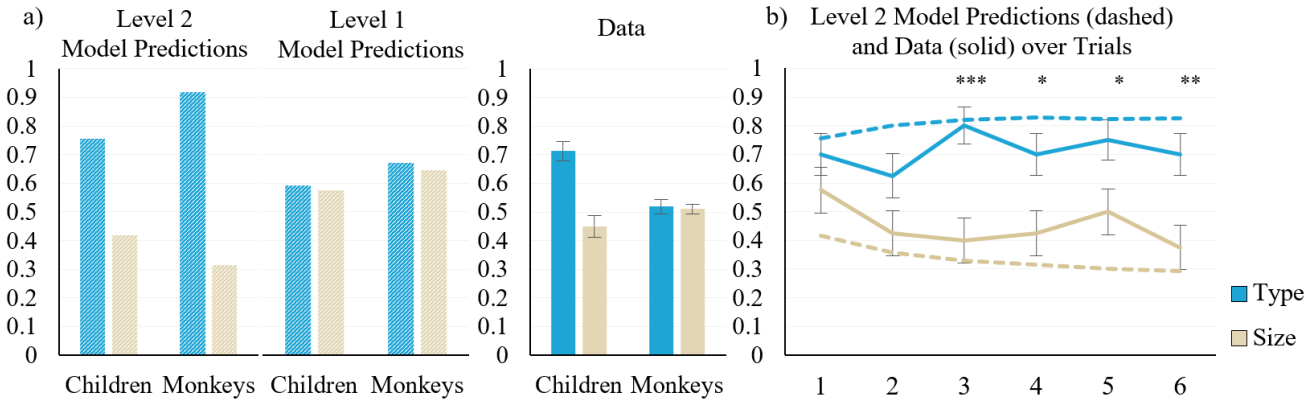

Figure 2: a) Model Predictions for a learner capable of Level 2 or Level 1 abstraction and empirical results (mean across trials ± SE) for the choice for the sample from the box with the small, high-valued example item for capuchin monkeys and children. Model predictions are shown for one trial with 3 evidence boxes. b) Children's level 2 model predictions and data (M ± SE) over the course of six trials. Significant differences between the size and the type condition are indicated.

both groups expressed a significant preference for the small, high-valued items over the big, low-value option (Capuchins: M=0.96, SD = 0.04, t(16) = 48.52, p <0.001; Children: (M = 0.92, SD = 0.21, t(57) = 15.68, p < 0.001). Further, no difference in performance was found between the choice presentation with crossed and straight hands.

Main Experiment. Monkeys were equally likely to choose the sample from the container with the small high-value example in both conditions (paired t(10)= 0.27, p = 0.79), and chose at chance level (12/24 trials) between the two hidden samples (type: M = 12.45, SD = 1.37, t(10) = 1.10, p = 0.30; size: M = 12.27, SD = 1.95, t(10) = 0.46, p = 0.65).

Unlike the preference testing, multiple monkeys expressed a bias regarding the side of their chosen reward sample or the side of the container (7/11 monkeys chose either a consistent hand-side or a consistent container-side in more than 80% of trials). They did not reach more frequently to the side of the small, high-valued sample (M = 0.52). There was no improvement from the first block to the second in either condition (type: M_{first} = 0.52, M_{second} = 0.52; size: M_{first} = 0.51, M_{second} = 0.51).

Children chose the sample from the container with the small, high-value example item more often in the type condition than the size condition, t(77.50) = -5.18, p < 0.001. When compared to chance (3/6 trials), only the choices in the type condition were significantly different (type: M = 4.28, SD = 1.41, t(39) = 5.70, p < 0.001; size: M = 2.7, SD = 1.30, t(39) = -1.45, p = 0.15).

The Level 2 models for both species predict a clear distinction between both conditions in the tendency to choose the item from the container with the small, high-value example item (Figure 2). Choice predictions are stronger for monkeys as the inferred utilities for low and high-value items based on their reward preferences are more extreme. When compared to the empirical data, the monkey's chance level performance is in stark contrast to the predictions of a model that learns overhypotheses, using item utilities inferred from the monkey's food preferences. For children the level 2 overhypothesis model predictions qualitatively fit the data well and the trajectory across trials shows a similar trend for both data and model predictions.

Discussion

As predicted by the model fit separately to their preference data, children made different choices in the size and type conditions despite the evidence from the test containers being the same in both cases, suggesting that they formed overhypotheses. However, their performance only differed significantly from chance in the type condition, which could suggest that they are only capable of forming abstract rules about certain reward properties. Alternatively, children might have a strong prior towards sorting by type, which is possibly more common in children's experience, or the two features might have had an unequal salience based on pre-existing preferences or the task description (see also Kemp et al., 2007 for discussion of the 'shape bias' in word learning). However, children did show a preference for larger items when presented with a simple choice in the preference test, suggesting they attended to this dimension. Interestingly, the overhypothesis model fit to children's preferences also predicted a smaller distinction from chance in the size condition, suggesting that this result may nonetheless be consistent with the overhypotheses. Future work could try to increase sample size or change utilities to differentiate lack of attention to the size dimension from a smaller predicted difference in utility between containers.

The monkeys' performance suggests that they were not able to form overhypotheses about the food distribution pattern across containers. Their failure on the second level of abstraction could be due to a failure to form abstractions about containers in general (Level 2 overhypotheses), or based on the inability to infer the content distributions of each evidence container (Level 1 overhypotheses) based on the sampled evidence. However, as with any negative result from a complex task, there could be other limiting factors, specifically the sampling procedure required sustained attention and inhibition skills which could be an impeding factor for the performance of monkeys (Tecwyn, Denison,

Messer, & Buchsbaum, 2017), a point we will return to in the general discussion.

Experiment 2: Abstraction within a container

To test the hypothesis that monkeys did not form Level 1 generalizations about the contents of the containers in experiment 1 (precluding generalization over containers – Level 2), we conducted a second experiment, with reduced task demands. Here, we presented subjects with only two containers from which we sampled four evidence items each. Now, the choice items were sampled directly from these containers, so that no generalization to new containers (Level 2) was required. However, participants would still have to form Level 1 generalizations to choose successfully.

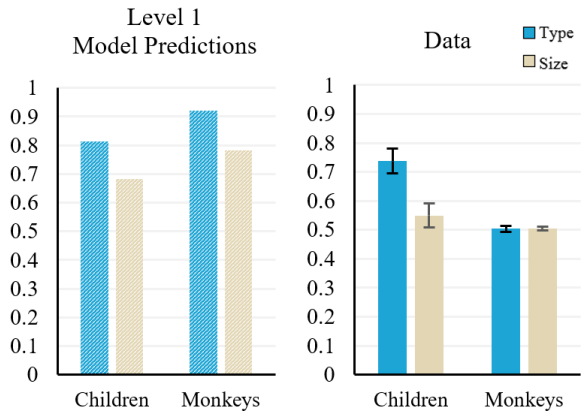

Figure 3: Model predictions (left) and empirical data (right, M ± SE) for the correct choices of children and monkeys in the type (high-value item) and size (large item) condition.

Methods

Participants. Participants were 47 4- to 5- year-old children recruited at two local museums in Toronto (M age = 5.0 yrs, 50% female, n= 24 in type condition, n= 23 size condition). Two additional children were excluded because they ended the game early or due to experimenter error. The total sample of capuchin monkeys (*Sapajus spp.*) consisted of 13 individuals. Ten had previously completed Experiment 1. Out of the 13 subjects, 11 participated in both conditions whereas two participated only in one of the conditions.

Design and Procedure. All sessions were video recorded. The procedure was similar to Experiment 1. This time only two containers were presented on the table and four items were sampled from each successively. Subsequently, the experimenter extracted the two choice items directly from these containers, kept them hidden in her hand and requested the participant to choose. In the size condition, the same four types of rewards, two low- and two high-value, were drawn from both containers in a randomized order whereby one container only yielded small (size 1) and the other one only big (size 5) items. The reward was identical to one of the four types previously drawn from the container. In the type condition, items of the same type in the sizes 1, 2, 4 and 5 were drawn from the container. Thereby one container offered only low-valued and the other only high-valued items. The reward was a randomly sized piece of the expected type for this container. Monkeys received 3 sessions of 8 trials each per condition with order of condition counterbalanced. Children received one session of 6 trials in a between-subject design beginning with a preference testing of two size and two type comparisons.

Results and Discussion

Children performed significantly above chance (3/6 trials) in the type condition ($M = 4.42$, $SD = 1.28$, $t(23) = 5.41$, $p < 0.001$) but not in the size condition ($M= 3.30$, $SD = 1.22$, $t(23) = 1.19$, $p = 0.25$). Monkeys performed at chance level (12/24 trials) in both conditions (type: $M = 12.09$, $SD = 0.94$, $t(10) = 0.32$, $p = 0.76$; size: $M = 12.09$, $SD = 0.54$, $t(10) = 0.56$, $p = 0.59$). The choice predictions of models based on the inferred feature distribution in each container (Level 1 abstraction) showed a clear tendency to choose the next item from the container with high-value items in the type condition and from the container with large items in the size condition. The strong type preferences of both species, lead to a greater predicted container preference in the type condition. Whereas the monkeys performed at chance level in both conditions, children's performance resembled the model prediction in both conditions, showing strong performance in the type condition whereas choices in the size condition were at chance. This suggests that children are able to form abstractions at both levels whereas capuchin monkeys in our study were unable to engage in any level of abstraction, though we emphasize that the reasons for this failure remain ambiguous (lack of ability or task demands).

General Discussion

We presented two studies testing abstraction, and the predictions of the Kemp et al. (2007) overhypothesis model, using a choice paradigm in children and capuchin monkeys. Across both experiments, none of the capuchin monkeys showed the pattern predicted for a learner capable of forming overhypotheses along the item size or type dimensions. In contrast, children treated the same evidence differently when they had previously experienced that items were sorted by size or type. Their performance was well characterized by a hierarchical Bayesian model, fit to their actual reward preferences. They showed a significant difference between conditions after just a few trials.

The model predictions based on capuchin's preferences support that the presented evidence was sufficient for the formation of overhypotheses, but the monkeys did not show this ability in this paradigm. The monkeys' results are in line with low success rates achieved after long training regimes in previous studies on abstract concept formation and analogical reasoning in capuchins (Flemming, 2011; Kennedy & Fragaszy, 2008; Truppa et al., 2011). We can also rule out some other possible explanations for their failure. Monkeys did not show a preference for the side exhibiting the small, high-value item (showing some

understanding of the procedure: they were not simply trying to acquire the samples). No individual monkey showed a difference between conditions, and the sample was sufficient to detect significant food preferences, suggesting that this was also not a sample size limitation.

Still, it remains possible that other tasks demands masked monkeys' abstract reasoning abilities. Monkeys and apes can infer a hidden item sampled from a clear population (Tecwyn et al., 2017; Eckert, Rakoczy, & Call, 2017; Rakoczy et al., 2014). However, apes' ability to make inferences about hidden populations based on visible samples (as in this study) was recently shown to be more limited (Eckert, Rakoczy, & Call, 2017). Future work will explore abstract reasoning with reduced task demands, e.g. by allowing the subjects to sample items themselves. Nevertheless, the approach taken here, in which subjects do not need to be trained to make arbitrary judgements about abstract relations but simply need to secure the best rewards, is a promising avenue for future research.

The findings from 4-5 year-olds are in line with infants' performance in causal learning and looking time procedures but stand in contrast to children's limited spontaneous use of abstract concepts in RMTS tasks (Hochmann et al., 2017), perhaps due to a reduced need for training. This approach could be extended to toddlers to bridge the gap across ontogeny.

In summary, we conducted the first direct test of the hierarchical Bayesian approach described by Kemp et al. (2007) in children and animals, and extended it to make choice predictions based on item utilities. We have shown that it is a promising model for how children are able to form generalizations from sparse evidence. We suggest that further application of computational models to empirical data of overhypothesis formation is desirable to understand its development over early childhood, and to further understanding possible species differences.

Acknowledgements

We thank Justine Biado, Kiah Caneira and Kay Otsubo. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [639072]). We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number 2016-05552]

References

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science*, *38*(2), 383-397.

Christie, S., Gentner, D., Call, J., & Haun, D. B. M. (2016). Sensitivity to relational similarity and object similarity in apes and children. *Current Biology*, *26*(4), 531-535.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge evidence from 9-month-old infants. *Psychological Science*.

Eckert, J., Rakoczy, H., & Call, J. (2017). Are great apes able to reason from multi-item samples to populations of food items?. *American journal of primatology*, *79*(10), e22693.

Flemming, T. M. (2011). Conceptual thresholds for same and different in old-(Macaca mulatta) and new-world (Cebus apella) monkeys. *Behavioural processes*, *86*(3), 316-322.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2003).Bayesian data analysis (2nd edn.). New York: Chapman & Hall.

Goodman, N. (1955). *Fact, fiction and forecast* (Vol. 74). Cambridge, MA: Harvard University Press.

Haun, D. B., & Call, J. (2009). Great apes' capacities to recognize relational similarity. *Cognition*, *110*(2), 147-159.

Hochmann, J. R., Tuerk, A. S., Sanborn, S., Zhu, R., Long, R., Dempster, M., & Carey, S. (2017). Children's representation of abstract relations in relational/array match-to-sample tasks. *Cognitive psychology*, *99*, 17-43.

Hochmann, J. R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, *177*, 49-57.

Hood, B. M. (2004). Is looking good enough or does it beggar belief?. *Developmental Science*, *7*(4), 415-417.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307-321.

Kennedy, E. H., & Fragaszy, D. M. (2008). Analogical reasoning in a capuchin monkey (Cebus apella). *Journal of Comparative Psychology*, *122*(2), 167.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*(6), 2797-2822.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109- ' 130.

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*(1), 60-68.

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology*, *25*(2), 256-260.

Tecwyn, E. C., Denison, S., Messer, E. J., & Buchsbaum, D. (2017). Intuitive probabilistic inference in capuchin monkeys. *Animal Cognition*, *20*(2), 243-256.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279-1285.

Truppa, V., Mortari, E. P., Garofoli, D., Privitera, S., & Visalberghi, E. (2011). Same/different concept learning by capuchin monkeys in matching-to-sample tasks. *PLoS One*, *6*(8), e23809.

Vonk, J. (2015). Corvid cognition: Something to crow about?. *Current Biology*, *25*(2), R69-R71.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, *25*(1), 161-169.

Wasserman, E. A., & Young, M. E. (2010). Same–different discrimination: The keel and backbone of thought and reasoning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(1), 3.