Cognitive Psychology 140 (2023) 101530

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Evidence for abstract representations in children but not capuchin monkeys

Elisa Felsche [a,b,*], Patience Stevens [c], Christoph J. Völter [d], Daphna Buchsbaum [e,1], Amanda M. Seed [a,1]

[a] School of Psychology and Neuroscience, University of St Andrews, Scotland
[b] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Germany
[c] Department of Psychology, Carnegie Mellon University, USA
[d] Messerli Research Institute, University of Veterinary Medicine Vienna, Medical University of Vienna and University of Vienna, Austria
[e] Department of Cognitive, Linguistic and Psychological Sciences, Brown University, USA

ARTICLE INFO

Keywords:
Overhypotheses
Abstraction
Generalization
Animal cognition
Computational modeling
Cognitive development

ABSTRACT

The use of abstract higher-level knowledge (also called 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. There is also mixed evidence for when this ability emerges over human development. Kemp et al. (2007) proposed a computational model of how overhypotheses could be learned from sparse examples. We provide the first direct test of this model: an ecologically valid paradigm for testing two species, capuchin monkeys (*Sapajus* spp.) and 4- to 5-year-old human children. We presented participants with sampled evidence from different containers which suggested that all containers held items of uniform type (type condition) or of uniform size (size condition). Subsequently, we presented two new test containers and an example item from each: a small, high-valued item and a large but low-valued item. Participants could then choose from which test container they would like to receive the next sample – the optimal choice was the container that yielded a large item in the size condition or a high-valued item in the type condition. We compared performance to a priori predictions made by models with and without the capacity to learn overhypotheses. Children's choices were consistent with the model predictions and thus suggest an ability for abstract knowledge formation in the preschool years, whereas monkeys performed at chance level.

1. Introduction

A primate foraging in a tree might find a first fruit that is a fig, continues with the search for food, and finds a second, a third and potentially a fourth fruit in the same tree, all figs. By now, the primate might learn "this tree grows figs" – a generalization at the first level of abstraction. If the animal moves on to another tree that bears some berries and subsequently picks a few nuts from a third tree,

* Corresponding author at: Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.
E-mail address: elisa_felsche@eva.mpg.de (E. Felsche).
[1] Equal contribution.

https://doi.org/10.1016/j.cogpsych.2022.101530
Received 10 December 2021; Received in revised form 2 October 2022; Accepted 23 November 2022
Available online 8 December 2022

it is exposed to the regularity that "trees carry a uniform fruit type". Learning this generalization at the second level of abstraction would make just one bite of a fruit from a new tree sufficient to decide whether to invest more time and energy foraging in this tree or if moving on to another food source might be preferable. The ability to detect recurring patterns and generalize information to new situations is a crucial tool that enables efficient knowledge gain. Abstractions represent generalized information that is common across situations; thus, they allow for a reduction of information by leaving out situation-specific details (Garlick, 2010). This makes them not reducible to concrete sensory input (Seed et al., 2011). Though human beings have a particular talent for abstraction, the ability to notice abstract regularities in the environment, make rapid inferences from limited data and transfer functional behavior across situations could be advantageous for other long-lived species in complex environments.

However, some argue that the ability for abstraction is central and unique to human intelligence (James, 1890) and fundamentally divides humans from other animals (Gentner, 2003, 2010; Locke, 1847; Penn et al., 2008). Others, challenging this view, argue that differences are of quality not of kind, such as the speed of acquisition, flexibility, or symbolic nature of the abstract knowledge (Katz et al., 2007; Premack, 2010; Seed et al., 2011). Some disagreement also exists regarding the ontogenetic onset of abstract thinking. The ability has long been thought to develop later in childhood, as preschool children often perform poorly in tasks measuring abstract relational reasoning (e.g., Christie & Gentner, 2010; Christie & Gentner, 2014; Hochmann et al., 2017; Richland et al., 2006). However, others suggest that the ability to make generalizable inferences from minimal data emerges much earlier and is a key component of human infants' and children's fast knowledge and skill acquisition, including word learning (e.g., Dewar & Xu, 2010; Xu & Tenenbaum, 2007b; Yin & Csibra, 2015). Thus, despite its high ecological relevance and decades of research, we are still unsure when this ability evolved and how it develops in human ontogeny. This is partly due to large methodological differences in the applied paradigms between age groups and the susceptibility of established tasks to prior biases and alternative explanations.

In this study, we introduce to the field of comparative psychology a hierarchical Bayesian computational approach (Kemp et al., 2007) that provides a formalism for how abstract knowledge can be learned from sparse data. Within this framework, we investigated whether preschool children and capuchin monkeys can infer abstract rules about the item distributions in containers when given only limited evidence and if they can use those abstract rules to guide behavior in new situations efficiently.

1.1. Background

The ability for abstraction in humans and non-human animals has most prominently examined the relations of "same" and "different". Infants in looking-based paradigms show an ability to detect these relations, for example, by showing habituation to pairs of matching or differing objects, even for displays involving new objects (for a review, see Hespos et al., 2021). Similarly, non-human animals can learn, given many trials, to engage in a specific response after seeing homogenous or heterogenous stimulus arrays (see Katz, Wright, & Bodily, 2007; Wasserman, Castro, & Fagot, 2017 for reviews).

However, the picture becomes more complicated, both in ontogenetic and phylogenetic comparisons, when participants are not just required to detect same/different relations but also to actively select an array with that same relation. In the relational matching to sample (RMTS) task, participants are presented with a sample stimulus pair (two identical or two different items) and then two choice pairs. They must select the pair with the matching abstract relation to the example (e.g., AA matches DD but not EF; BC matches EF and not DD). In some studies, children only show reliable success by age 5, with further improvement into adulthood (e.g., Christie & Gentner, 2010, 2014; Hochmann et al., 2017; Kotovsky & Gentner, 1996; Thibaut et al., 2008). However, small task modifications like verbal scaffolding (e.g., labeling the example object pair with a novel word), presenting multiple examples, or using larger item arrays enable children to succeed at younger ages (Christie & Gentner, 2010, 2014; Gentner et al., 2011; Hochmann et al., 2017; Kotovsky & Gentner, 1996). For example, in Goddu, Lombrozo and Gopnik (2020), 3- and 4-year-olds succeeded in a causally-framed RMTS task. Here, children saw two examples of how a wizard transforms objects (e.g., growing an apple and a dog from small to big). Subsequently, they chose the correct relational pair (a small & a big cube) over a conflicting object match (e.g., round and flattened apple) more often than a control group that experienced the classic RMTS procedure. In a blicket detector task, 18- to 30-month-old toddlers correctly selected object pairs following evidence that their relation was causally relevant to activate a machine (Walker et al., 2016; Walker & Gopnik, 2014, 2017). Interestingly 30–48 month-olds did not succeed, perhaps reflecting an increasing prior expectation that object identities, not object relations, are causally relevant for activating a machine.

Only a few non-human species master the RMTS task, usually achieving success gradually after lengthy training periods, with better performance on multi-stimulus arrays rather than stimulus pairs, and while showing some reduction in performance in transfer compared to training trials (Katz, Wright, & Bodily, 2007; Wasserman, Castro, & Fagot, 2017; Wasserman & Young, 2010). For example, out of five capuchin monkeys in Truppa et al. (2011), only one solved 2 and 4-item array RMTS trials after experiencing over 20,000 trials. Some have suggested that success in same/different and RMTS tasks, especially under these circumstances, can result solely or at least to some degree by matching the perceptual variance within item displays, which does not require a grasp of the abstract relation (Penn, Holyoak, & Povinelli, 2008; Vonk, 2015; Wasserman, Castro, & Fagot, 2017). Conversely, some argue that failure on RMTS tasks need not result from a lack of abstract reasoning ability per se. Instead, it could be caused by inductive biases, like English-speaking preschoolers' focus on individual stimuli (e.g., Carstensen & Walker, 2017; Christie et al., 2020; Christie & Gentner, 2010; Walker et al., 2016), which prevents them from detecting the correct relational solution (Kroupin & Carey, 2021). Kroupin and Carey (2021) argue that the slight task modifications that help children and non-human animals succeed, like the presentation of labels and multiple examples or previous variable matching to sample training, serve to shift those biases rather than suddenly enabling abstract thinking. Given the importance of prior experience and knowledge, RMTS tasks in their current form, despite their prominent use, might not be ideal for assessing species differences and developmental trajectories in the ability to form and use abstract concepts.

Outside of RMTS tasks, the evidence is also mixed and debated. Some positive evidence for non-human great apes' (Christie et al., 2016; Flemming & Kennedy, 2011; Haun & Call, 2009) and capuchin monkeys' (Kennedy & Fragaszy, 2008) abstract reasoning ability comes from food searching tasks that require primates to reason based on spatial relations (e.g., "top" or "middle" cup) or relative sizes (e.g., largest cup). In these tasks, participants see a food item hidden in one array of cups. They have to find their reward in a second array of objects in the corresponding location. Positive evidence has also been found from studies using other naturally-occurring relations, such as biological categories (e.g., humans vs other animals; Benard et al., 2006; Lazareva et al., 2004; Tanaka, 2001; Vonk et al., 2013) or physical concepts such as connection, solidity and support (e.g., Mayer et al., 2014; Seed et al., 2011; Seed et al., 2009). However, whether or not these tasks are solved using abstract concepts has been a matter of debate due to issues such as long training periods, alternative explanations for successful task performance based on surface perceptual features, and/or success in only a small minority of the sample (e.g., Brooks et al., 2013; Cook, Wright, & Drachman, 2013; Penn, Holyoak, & Povinelli, 2008; Povinelli & Penn, 2011. Infants and some species of birds and non-human primates have successfully distinguished between abstract stimulus patterns like ABA and AAB sequences (Marcus et al., 1999; Sonnweber et al., 2015; Spierings & Ten Cate, 2016). However, sample sizes were also limited here, and for some findings, lower-level strategies have been proposed as alternative explanations for seemingly abstract generalization (Corballis, 2009; Neiworth et al., 2017; Ravignani et al., 2015; Spierings & Ten Cate, 2016). Further, some negative findings may be due to ecologically ill-matched stimuli in comparative studies (Ravignani et al., 2019).

In summary, there is no agreement as to whether abstract knowledge formation is an evolutionary primitive, shared with other species and emerging early in human development, or a recently-evolved and late-developing skill. Interpretation of existing evidence on this topic is complicated by considerable methodological differences between the tasks used across ages and species. Whereas non-verbal habituation and looking-time paradigms are primarily used with infants, the literature on older children's abstraction abilities is dominated by verbal relational reasoning tasks. Such tasks often impose additional challenges and are prone to a learned bias for individual objects. Further, the current comparative literature has focused on paradigms that disentangle abstraction from purely perceptual strategies. They often fail to do so, especially when extensive training regimes are provided. However, abstractions cannot be achieved in isolation but depend on perceptual processing in the first place (Wasserman, Castro, & Fagot, 2017). Thus, a more dynamic approach to abstract reasoning that accepts the influence of perceptual processes is desirable, in which both surface features and more abstract relations are learned simultaneously rather than being in opposition to one another.

Such an alternative approach was provided by Goodman (1955), who introduced the term 'overhypothesis' to describe the process by which abstract generalizations can be formed that inform inferences about specific new examples (Kemp et al., 2007). Analogous to the tree example from the start of the paper, Goodman (1955) provides the idea of marble-filled bags to illustrate the concept of overhypotheses. After seeing just a few randomly sampled marbles from a few bags (e.g., four green marbles from the first, four white marbles taken from the second, and four red ones sampled from the third), we can infer the colour of the rest of the marbles in each bag (Level 1 abstraction) as well as the overhypothesis that 'bags contain marbles that are uniform in colour' (Level 2 abstraction). This emerging overhypothesis constrains the hypothesis space at lower levels of abstraction. Before opening a new bag, we predict that all of them will have the same colour rather than a mix of colours, even though we are naïve about the exact colour type. If we retrieve a single blue marble from the new bag, we can predict that all of the other marbles in this bag will be blue using the overhypothesis.

Many familiar abstract concepts can be thought of as overhypotheses, perhaps to some degree the two terms are synonymous. The concept "mammals" can be described as an overhypothesis that "all mammals share a set of features" (e.g., mammary glands, hair, mostly born alive). The possession of this overhypothesis, built from experience, can support inferences about new exemplars that match the category features: a newly-encountered hairy quadruped might be expected to give birth to live young. At the same time, a scaled cold-blooded individual might be less so. Moreover, the overhypothesis 'animal' at a higher level of abstraction licenses inferences about the need for ingestion and respiration. Similarly, the concept "same" is built from specific examples like two blue circles, five red triangles or 20 bananas. The central requirement for forming an abstract concept or an overhypothesis is the detection of commonalities between examples and the generalization of observed relational patterns to other item arrays. However, the overhypotheses framework usefully emphasizes the multiple levels and forms of abstraction that might simultaneously be detected over populations of items, and the role of abstract knowledge in helping us to infer or predict something about future novel instances.

Goodman's marble example illustrates how abstract categories like same and different are formed based on the seen evidence and applied to novel situations, while tasks such as the RMTS probe the possession of abstract knowledge by assuming that these are already in place. Given the evidence discussed above that groups other than human adults may preferentially match on other dimensions (e.g., based on object identity, colour composition or number of corners), the overhypothesis approach in which both the evidence and the predictions are quantified could be a fruitful way to develop paradigms in which the role of prior experience can be taken into account.

Kemp and colleagues (2007) proposed a probabilistic hierarchical Bayesian model as a normative model of how the overhypothesis in the Goodman scenario can be learned from such a limited amount of sampled data. In general, probabilistic hierarchical Bayesian models suggest how overhypotheses can be acquired non-linguistically from sparse data, and used to make wide-ranging predictions (Kemp et al., 2007; Lucas & Griffiths, 2010; Tenenbaum & Griffiths, 2001; Tenenbaum et al., 2006; Tenenbaum et al., 2011). The models take in the learner's observations and use them to infer the probability of overhypotheses at multiple levels of abstraction simultaneously, explaining the current data (e.g. seen marbles) while both considering and forming prior assumptions about how the world generated these data (e.g. uniformity of marbles in bags). Judgments are influenced by both concrete observations and higher-order concepts, explaining the formation of abstract concepts via the learner's perceptual input. Different aspects of the Bayesian formalization, such as the influence of prior knowledge, the simultaneous consideration of multiple hypotheses, and the systematic shift of their probability when encountering new evidence, have been shown to play a role in children's inferences (see Gopnik & Wellman, 2012 for a review). Bayesian models have been used to capture a variety of inductive learning phenomena like the

acquisition of new category labels (Xu & Tenenbaum, 2007a, 2007b), the influence of prior biases when inferring causal principles (Lucas et al., 2014), how the social context shapes causal learning (Goodman, Baker, & Tenenbaum, 2009) and goal inference when observing intentional agents (Ullman et al., 2009). For example, in a blicket detector design, children and adults had to judge the causal power of objects to turn on a machine after seeing evidence for individual objects and object pairs (Griffiths et al., 2011). A Bayesian model captured the participant's judgments and correctly predicted the outcome of a change in the prior knowledge (causal power is rare or common among objects) and whether the causal connection was deterministic or probabilistic.

In a looking-paradigm inspired by the Goodman marble task, Dewar and Xu (2010) presented 9-month-olds with evidence supporting the overhypothesis that containers are filled with objects of the same shape but different colours. In a test situation, infants who had seen this evidence looked longer when two differently shaped objects were drawn from a new container (contradicting the overhypothesis) than when two objects of the same shape were sampled. Infants who had not seen the evidence showed no significant difference in looking time to the two outcomes. Thus, the authors concluded that already at the age of 9 months, infants can form inductive overhypotheses from limited evidence. This finding is in line with other looking-time studies showing that infants can extract and generalize the relation same (but possibly not the relation different; Hochmann, Carey, & Mehler, 2018) after repeated exposure (see Hespos et al., 2021 for a review).

Given the discrepancies between the positive findings from looking-time studies in infants, and active measures in preschool children and non-human animals in which feedback is often conflated over multiple dimensions, investigating whether preschool children and other primate species can spontaneously form overhypotheses in this way and use them to make predictions would bridge a methodological gap.

In this study, we introduce a novel task based on the original idea of overhypothesis formation by Goodman (1955) that can be used across species to examine abstract knowledge formation in an ecologically valid context without extensive training or explanation. We adapted Goodman's thought experiment, in which bags of marbles can be either uniform or mixed in color, to create an active choice paradigm without training suitable for older preschool children and capuchin monkeys. Importantly this task design allows us to test the a priori predictions of a theoretical computational model, for how limited data can be sufficient for overhypothesis formation in this same task (extending the model of Kemp et al., 2007).

2. Experiment 1: Abstraction across containers

We presented a conceptually similar task to both children and monkeys and to our normative computational model of overhypothesis formation. In the task, the experimenter samples evidence items from three containers (food items for monkeys and sticker strips for children). The sampled evidence either supported the overhypothesis that rewards are sorted into containers by their type (e.g., drawing four banana pieces of various sizes from the first container, four apple pieces from the second container, etc.) or that they are sorted by their size (e.g., four small items of different types from the first container, four large items from the second container, etc.). 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. Participants then chose between two covert samples from these new containers. A differential choice between conditions—namely, selecting the next sample from A to obtain a high-value option in the type condition but choosing the next sample from B to receive a large item in the size condition—would reflect sensitivity to the overhypotheses governing object sorting. Because participants in the two conditions are presented with identical stimuli at test, and have to use their prior learning to select between covered samples from new item populations, the lower-level explanation for success in RMTS tasks (comparing the perceptual variability between item arrays, e.g., Penn et al., 2008) cannot readily be applied to success in our study.

In the following, we first describe the methods for primates and children. Our goal was not quantitative comparison, and there were some differences in the instantiation to adapt to the different testing constraints, but by conducting a conceptually similar task we aimed to draw qualitative comparisons. We next present the model implementation details and *a priori* model predictions for the choices that an idealized learner of overhypotheses, with the same item preferences as either the children or the monkeys, should make in our task.

2.1. Method

2.1.1. Participants

We tested 80 4- to 5- year-old children (40 female, $M_{age} = 4.9 \pm 0.6$ years (calculated based on the birthdates of 67 children; for 12 5-year-olds and one 4-year-old no birthdates were provided), recruited at two local museums in Toronto, Canada. The participants' age and gender were counterbalanced across conditions (type condition: $n = 40$; $M_{age} = 4.9 \pm 0.6$ years, 20 female, 10 missing birthdates; size condition: $n = 40$, $M_{age} = 4.9 \pm 0.6$ years, 20 female, 3 missing birthdates). Eight additional children were excluded from the analysis because they ended the game early ($n = 5$) or due to experimenter error ($n = 3$). The study was developed and conducted in accordance with ethical guidelines. It was approved by the ethics committee of the School of Psychology and Neuroscience at the University of St Andrews and by the Institutional Research Ethics Board for Human Subjects at the University of Toronto. The parents of all children who participated had given prior consent for their participation. Further, we explained to the children that they could stop participating at any point and asked them multiple times throughout the procedure if they would like to proceed with the experiment.

Seventeen brown capuchin monkeys (*Sapajus* spp., $M_{age} = 6.5$ years, 5 female) completed an initial food preference testing. Due to motivation decline only 11 of these monkeys finished the main study in the available time and were included in the data analysis (see SI for details). The monkeys are housed at the Living Links Research Center for Human Evolution in Edinburgh Zoo. They live in two large

social groups and cohabit their enclosures with squirrel monkeys (*Saimiri sciureus*), with whom they also coexist in their natural habitat. All individuals were born in captivity and mother-reared, except for one wild-born and hand-reared monkey. Participation was entirely voluntary, and the test sessions did not last longer than 15 min. We split the sample into two groups while counterbalancing the monkeys' sex and age and randomly assigned the groups to a starting condition (within-subject design). The monkeys had access to the test cubicles at two timeslots during the day (morning and afternoon). Thus within a day, up to two sessions could be conducted. The experimental protocol and study design for the monkeys were approved by the ethics committee of the School of Psychology and Neuroscience at the University of St Andrews.

2.1.2. Procedure

Reward Preference Testing. Before the main experiment, we conducted preference testing (see SI for details) to ensure that participants preferred bigger over smaller (size comparisons) and same-sized high- over low-value rewards (type comparisons). As a baseline for later test comparisons, small, high-value items were also compared to large, low-value items (mixed comparisons). Reward items were presented in a covered forced-choice procedure, where the participant first saw the reward items on the experimenter's palms and then had to choose between her closed fists. Our computational model uses data from this preference testing to predict the species-specific choice behavior based on their inferred item preferences.

Monkeys. We presented monkeys with nine different types of food items (divided into three categories: high-, medium- and low-value) in five item sizes. Monkeys received nine kinds of size comparisons (size 5 vs size 1), one for each food type. Further, we conducted six kinds of type and mixed comparisons, respectively, where each of the three high-value items was compared to two low-value items. Finally, the least liked high-value item (grape) was compared to all three medium-valued items to ensure a clear preference. Monkeys received ten trials for each of the 24 comparisons, presented over 24 sessions. The experimenter crossed her hands in half the trials to match the later test procedure.

Children. Rewards for children were stickers picturing either animals (high-value) or simple shapes (low-value; see Fig. 5). 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 tasked with filling in a blank zoo map (see SI for further details) with as many animal stickers as possible, making those more valuable than shape stickers. Afterward, the children received a warm-up of three trials in which they were familiarized with the closed-hand choice procedure and the task by choosing a hand with a reward item over an empty hand. Due to the constraints of museum testing, children were subsequently 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, we used novel stickers for the main experiment to ensure high motivation and asked children to find a lot of animals for a new, blank zoo map. As children participated only in one session and thus had much less opportunity to get used to this procedure, we only presented them with choices between uncrossed hands.

Main Experiment. All sessions of the reward preference testing and the main experiment were video recorded. For both species, the procedure in each trial was very similar (see SI for details on the used materials). The experimenter successively sampled four items from each of three evidence containers into transparent cups (monkeys) or onto metal frames (children), always starting on the same 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 Figs. 1 and 2). After four items were sampled from a container, it was covered with a lid so that it could not be associated with subsequent food items. 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 sampled evidence still in view of

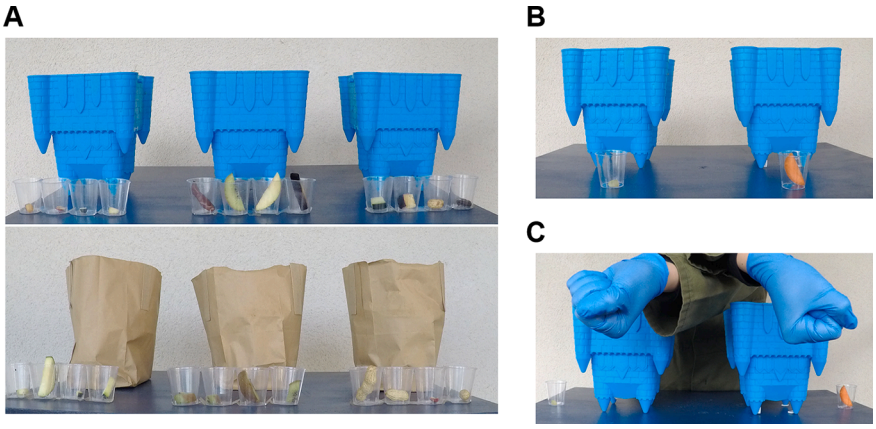

Fig. 1. Pictures of materials as used with monkeys. A: End of evidence presentation for a main experiment trial. The upper photograph shows the entire evidence for a trial in the size condition (order: small, large, and medium-sized items), the lower picture shows a completed evidence phase for a trial in the type condition (here: zucchini, kiwi, and peanut items). B: Presentation of the two test boxes and the respective small, high-value (here a piece of grape) and large, low-value (here a piece of carrot) example food samples. C: Choice situation with example items moved to the side and presentation of the choice items kept hidden in the experimenter's hands (here, the experimenter's hands are crossed as in half the trials).

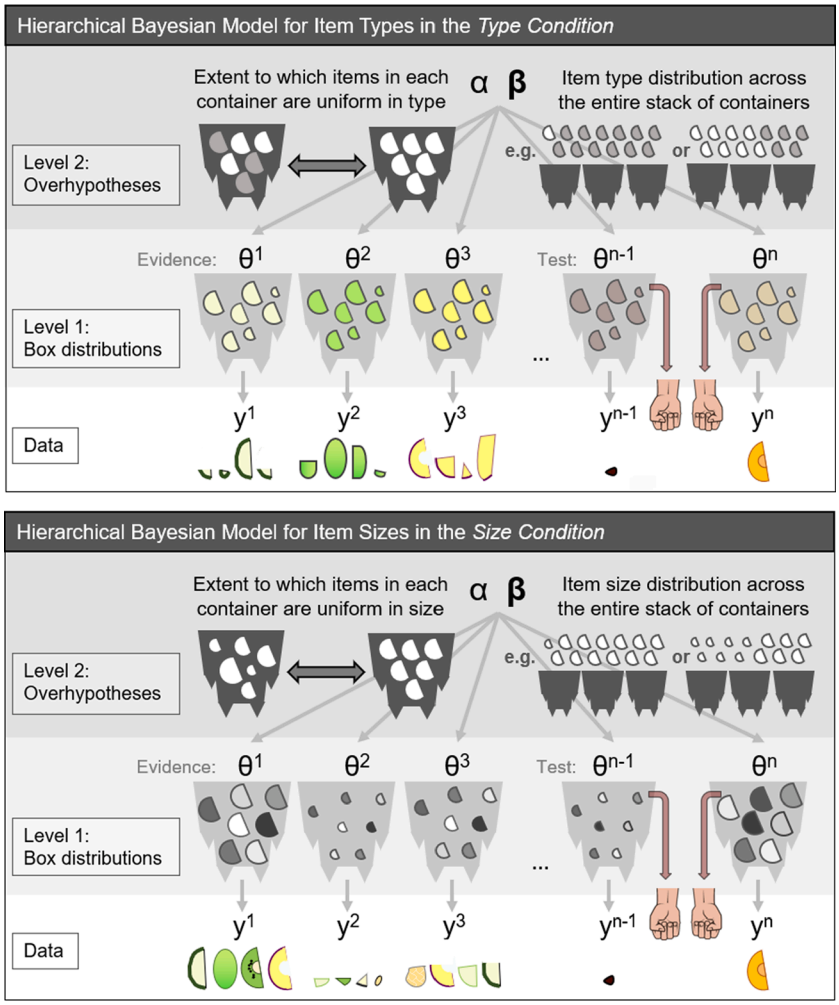

Fig. 2. Hierarchical Bayesian model of overhypothesis formation. For illustration purposes we displayed overhypotheses over item type distributions for the type condition (top panel) and overhypotheses over item size distributions for the size condition (bottom panel). However, the model generated posterior distributions for both item features (size and type) in both conditions. 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 (**θi**, Level 1 abstraction) are constrained by overhypotheses at Level 2 and, in turn, constrain the items ***yi*** sampled from that container. Ultimately the inferred item distributions in the test containers are of interest as from those, the experimenter samples the hidden choice items.

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 (sides 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 toward the participants so that they could indicate their choice by reaching toward one of the hands. Participants were rewarded with the chosen item. Reward items were selected to be in line with the condition-specific overhypothesis (i.e., of the expected type or size), at least of medium value in the size condition, and otherwise randomly sampled (see SI for further details on the counterbalancing). The non-chosen reward item was put away without being seen by the participant. All other sampled items were also removed so that participants would not associate the samples with the subsequent trial.

For monkeys, the experimenter crossed her hands in half of the trials (a procedure they were familiarized with during preference testing) to ensure they tracked the hidden sample in the experimenter's hand and were not just pointing towards the side of the preferred sampled example item (Tecwyn et al., 2017). For children, as in the preference testing, hands were never crossed. In comparison to the monkeys, children's pointing was not restricted by a choice panel; thus, they could indicate a specific hand rather than only a side. To better compare the two species, the children received no explicit instruction concerning the abstract rules governing the reward distribution. That is, despite the explanation of the goal to fill in a blank zoo map in the beginning (see SI for further details), we only prompted them to choose in the first trials of the preference testing with remarks like "Which one?". When we presented the test boxes for the first time, the experimenter said, "Let's see what is in this box" before starting to sample. After the last

choice of the child (to ensure we would not influence any choices), before opening the chosen hand, the experimenter shook it and asked the child, "What do you think is in this hand?". In case of no answer, the experimenter further proposed a choice question. In the type condition, she asked for example, "Do you think it is a lion or a square?" always referring to the two types already present in the example items sampled earlier from the test boxes. In the size condition, it was asked, "Do you think it is only one sticker or a long strip of stickers?". With these questions we were aiming for a deeper insight into the motivations behind children's choices.

2.1.3. Design

We contrasted two levels of one variable in a within-subjects design for monkeys and a between-subjects design for children. In the size condition, we presented evidence consistent with the overhypothesis "containers are filled with objects of the same size". In the type condition, the presented evidence followed the overhypothesis "containers are filled with objects of the same type". Due to the small sample size, monkeys experienced both conditions, type and size, in an ABAB design. Here blocks of four sessions were alternated so that each monkey participated in eight sessions of the type condition and eight sessions of the size condition. Each of the 16 sessions comprised three trials (for a total of 24 trials per condition, 12 trials in each block). Eight monkeys started with the type, and nine began with the size condition. Each monkey was presented with two different kinds of containers, bags, and boxes (counterbalanced regarding order and condition), so that each overhypothesis could be tied to a specific sort of container (e.g., the type condition was presented in bags and the size condition in boxes or vice versa).

Children were tested in a between-subjects design to allow us to test them in a single session in a science museum. Thus, they were only presented with one container type (boxes). Children received one session of up to 9 preference testing trials (including familiarization) and 6 trials of the main experiment, adding up to about 15 min. The monkeys' extensive preference testing was presented in multiple sessions, separate from and before the main experiment (see SI for details).

2.2. Computational model

While probabilistic hierarchical Bayesian models like that by Kemp et al. (2007) have successfully characterized existing findings of children's rapid early learning, 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 learners should choose with new empirical data from two species. We adapted the model to group-specific factors like the relative utilities of the different reward types. To evaluate the participants' ability for abstract knowledge formation, we compared their performance to models differing in their capability for abstraction at various levels.

Model Overview. Fig. 2 provides an overview of our task and the probabilistic hierarchical Bayesian model adapted from Kemp et al., 2007. As in our task, 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 the first level of abstract knowledge (Level 1), describing the kinds of items likely to be found in this specific container. The model also represents a more abstract level of knowledge (Level 2), which describes the probability distribution over containers—the extent to which containers tend to be mixed, uniform, or somewhere in between, and the distribution of features across containers (e.g., there are primarily big items and only a few small items). Using this hierarchical structure, the model captures how specific observations of samples from individual containers can be used to infer distribution parameters at multiple levels of abstraction simultaneously. While this model captures predictions about the distribution of items within and across containers, it does not predict how a learner should choose which container to receive a reward from in order to maximize their utility. Thus, we added a rational choice rule, allowing us to compare the model's predictions for which test container (A or B) learners should choose from to maximize their reward outcome, given the items they observed being sampled and their own utility over items. We infer the relative utilities of the different reward types for each species based on the participants' choices in preference testing, following the model for inferring item preferences from choices developed by Lucas et al., 2014. Including inferred item utilities in our choice rule is an important sanity check for our task. While intuitively, it may seem obvious that a rational choice would be for the container that is more likely to yield a high-value item, the effect size will be driven by how much more valuable a high versus a low-value item is.

Learning Overhypotheses. As in Kemp et al., 2007 we use a Dirichlet-multinomial model (Gelman, 2003) to describe how hypotheses at different levels of abstraction are derived from the sampled evidence and, in turn, also predict future samples. 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. Each container's distribution, representing the first level of abstraction, is described by a multinomial function (θ_d^i) from which the item types (sizes) are sampled $y_d^i \sim$ Multinomial (θ_d^i). Each container's distribution over item types (sizes), θ_d^i, is in turn determined by an overhypothesis at the second level of abstraction, thus, sampled from a Dirichlet distribution, parameterized by a scalar α_d and a vector $\boldsymbol{\beta_d}$, $\theta_d^i \sim$ Dirichlet(α_d, $\boldsymbol{\beta}_d$). These hyperparameters (α and $\boldsymbol{\beta}$) characterize the overhypothesis across containers. α_d parameterizes the extent to which items in each container are uniform in type (size), and in theory, could range from containers only producing samples of one item type (size) to a complete mixture of all available item types. $\boldsymbol{\beta_d}$ represents the type (size) distribution across the entire set of containers. For example, depending on the seen evidence, $\boldsymbol{\beta}$ could describe an item pool of 60 % apples and 40 % bananas, with 70 % big items and 30 % small items or, as the evidence suggests in our case (for the monkeys), an item pool composed of 9 food types in 5 sizes, all present in equal parts. The hyperparameter α_d is sampled from an exponential distribution, $\alpha_d \sim$ Exponential(1), making uniform item distributions in containers more likely than complete mixtures. The hyperparameter $\boldsymbol{\beta_d}$ is sampled from a symmetric Dirichlet distribution, $\boldsymbol{\beta_d} \sim$ Dirichlet(1), to not a priori bias the model to assume an unequal

distribution of item types (or sizes) in the item pool. When the model is now presented with evidence, those hyperparameters are flexibly updated and will reflect whatever the evidence suggests, e.g., 50 %, 73 %, 91 % or 100 % of items in a container are usually the same. Thus, the model is naïve to the experimenter-determined conditions of same and different but will learn them given sufficient evidence. To model overhypothesis formation, we infer $p(Y_d)$ (referred to as $p(Y)$ for simplicity below), the posterior distribution over (α, **β**), given the observed items $\boldsymbol{y^i}$, drawn from the *N* evidence containers,

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

estimated using the Metropolis-Hastings algorithm. Here we used 5 chains with 2000 samples and a burn in of 1000. Intuitively, an overhypothesis that containers are mostly uniform in type (size) would be represented as a posterior distribution over α shifted towards smaller values, while an overhypothesis that containers are mixed would be shifted towards larger values of α (see SI for how the posterior distributions of α and **β** values change given different amounts of evidence).

Predicting the Content of the Test Buckets. To estimate the subject's final choice, we first need to predict the type (size) of the next (unseen) sample *j* 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 (see the previous paragraph). The probability for this next sample to be of a certain type (and size) equals the predicted proportion of this type (size) in the new container. For a Dirichlet-Multinomial distribution, $p(y_j^{i+1}|y_{-j}^{i+1},\alpha,\beta)$, the posterior predictive distribution for the type (size) of the next item in the container, given the previously seen items from this container and the hyperparameters $\alpha, \boldsymbol{\beta}$, has a simple closed form solution. Marginalizing over $p(\alpha,\beta|Y)$, the posterior distribution over possible values of α and **β**, estimated from the evidence containers give us,

$$p(y_j^{i+1}|y_{-j}^{i+1}) = \iint p(y_j^{i+1}|y_{-j}^{i+1},\alpha,\beta)p(\alpha,\beta|Y)d\alpha,d\beta \tag{2}$$

approximated by averaging $p(y_j^{i+1}|y_{-j}^{i+1},\alpha,\beta)$ across sampled values of $p(\alpha,\beta|Y)$ (see SI for predictions of the item distributions in both test containers for different amounts of evidence).

Predicting the Choice of the 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 × 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 ×, is $u_x = s_x \bullet \delta_{t_x}$, where δ_{t_x} is the learner's utility for one unit of item type t_x. For example, if a high-value item has a baseline utility of 2, size 4 of this item would have a utility of 8. 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 (Lucas et al., 2014), we assume that learners become exponentially more likely to choose a container *i* as its expected utility increases (Luce-Shepard choice rule; Luce, 1959; Shepard, 1957)

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

Inferring reward utilities. To compute the relative utilities of the different reward items, we used the previously performed preference trials comparing different types of the same size, different sizes of the same type, and small items of high value with large items of low value (only closed hands version, see Methods). For simplicity, we only included the categorical item types high-, medium- and low-value for monkeys and high- and low-value for children. 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 $p(c) \propto p(u)p(u)$, estimated using the Metropolis-Hastings algorithm. For example, 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 10,000 samples and a burn in of 500 (see SI for how choice predictions differ giving varying strengths in preferences).

Reduced Level 1 Model. To achieve a richer data-to-model comparison and capture potential intermediate developmental or evolutionary stages, we formulated in addition to a learner capable of Level 1 and Level 2 abstraction also a learner only capable of Level 1 abstraction. For this lesioned model, we set α and β to fixed prior values, so that they were not updated by the evidence, however, θ_i values were still updated by the evidence. In other words, this model can update its representation of what is in a container from previous items sampled from that container, but it cannot update its representation of what containers are like in general. Thus, as all test samples were either high-value, size one or low-value, size five, predictions for the test buckets were created only considering high- and low-value types of the sizes one and five.

Model Predictions. We inferred strong preferences for high- vs low-value items for both species (value difference between high- and low-valued items: children: Δ0.88; monkeys: Δ1.20), confirming our empirical findings that these items vary significantly in utility. 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 evidence containers; Fig. 3A). Choice predictions differ between species for monkeys as the inferred utilities for low and high-value items based on their reward preferences are more extreme. After seeing up to 6 sets of evidence containers, predictions across subsequent trials for children get asymptotically more extreme (Fig. 3C). These predictions confirm that the data provided to participants in our experiment is sufficient to induce a strong overhypothesis and differential choice behavior in a learner capable of representing overhypotheses.

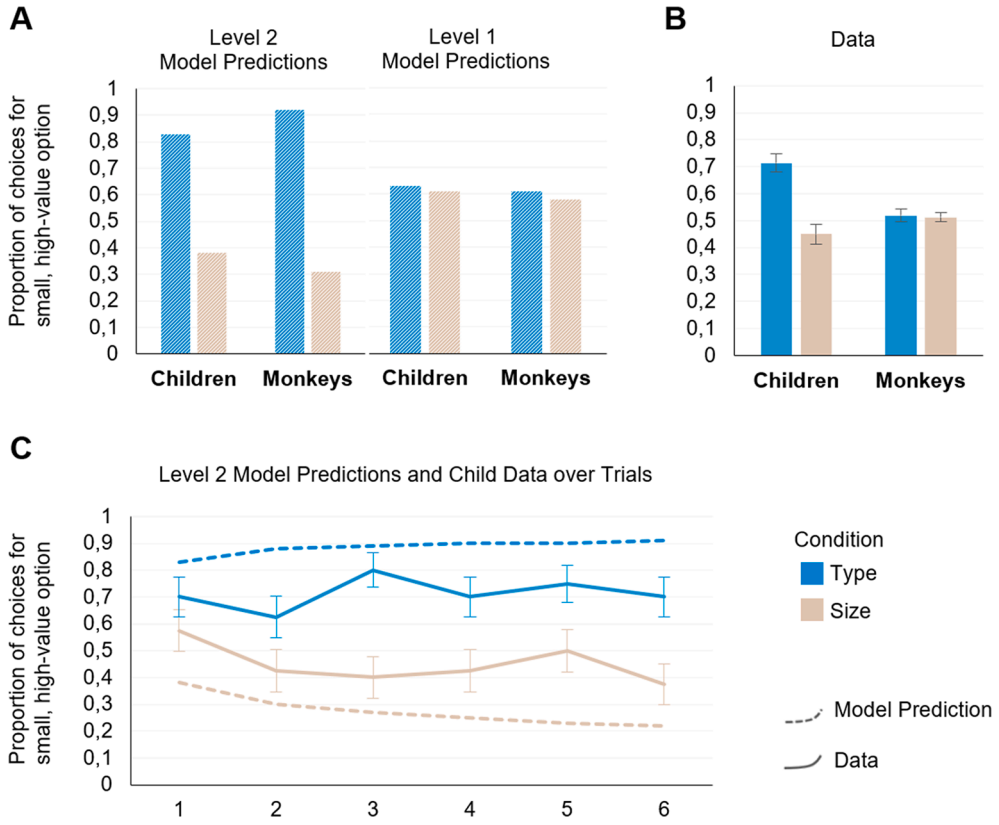

Fig. 3. Results and model predictions for the main experiment. A: Model Predictions for a learner capable of Level 2 or Level 1 abstraction for the choice for the sample from the box with the small, high-valued example item for capuchin monkeys and children. Model predictions are shown for one trial with three evidence boxes. B: Empirical choices for the hidden sample from the container with the small, high-valued example item for both species and conditions (mean across trials ± SE). C: Children's Level 2 model predictions and data (M ± SE) over all six trials.

These predictions contrast with those of the lesioned model capable of only Level 1 abstraction, which instead of generalizing from the evidence containers to the test containers, simply learns about the test containers only from the test samples. For this model, the test container with the small, high-value item is the slightly preferred choice independent of condition (58–61 % in monkeys and 61–63 % in children, the small descriptive difference between the size and type condition stems from the circumstance that we used three sizes in the size condtion (1,2 and 3) but 4 sizes in the type condition (1,2,4 and 5)). This bias resembles both species' preference for small, high-valued over large, low-valued items, as shown in the mixed comparisons of the preference testing. However, the tendency to choose the next sample from the container with the small, high-values example item is less extreme than in the preference testing, as the single example items in the test situation only have some predictive power regarding the item distributions in containers.

Finally, we consider a learner with no abstraction abilities. Such a learner cannot learn about the distribution of the items in a container from previous samples. We predict that such a learner makes random choices independent of any evidence. Thus, the choice rates for the sample from the container with the small, high-valued example item are expected to be at 50 % in both conditions.

2.3. Scoring and analysis

Preference Testing. In both species, we scored for every trial whether the participants chose the high-value item (based on our a priori categorization, see SI) in the type comparisons, the large item in the size comparison and the high-value item in the mixed comparisons. We compared performance to chance separately for each kind of comparison using two-sided *t*-tests.

Main Experiment. All analyses were performed in R (version 4.0.3, R Core Team, 2020). For both species, we measured in each trial whether participants chose the hidden sample from the container with the small, high-valued example item. Whereas in the type condition, subjects are expected to choose this option (to receive a high-valued item), in the size condition, they should choose the alternative sample (to receive a large item). A second coder, blind to the purpose of the study, coded choices for 20 % of all monkey sessions and 25 % of all child sessions, randomly chosen. Interrater agreement was excellent (monkeys: Cohen's kappa = 0.98, $p < 0.001$; children: Cohen's kappa = 0.94, $p < 0.001$). As for the food preference testing, all means and standard deviations are reported as proportions of 1 to ease comparability between species.

Capuchin Monkeys. We used a *t*-test for dependent samples to compare the monkeys' choices between conditions, as their data were normally distributed (Shapiro-Wilk test). We also used one-sample *t*-tests to test their scores in both conditions against chance.

Children. We used a *t*-test for independent samples to analyze children's choices between conditions. Further, we investigated whether children's performance became more condition-specific over trials by testing the interaction of condition and trial number using generalized linear mixed models (GLMM) with a binomial error structure. The models were fitted using the function glmer of the lme4 package (Bates et al., 2014). We used likelihood ratio tests to assess whether the interaction of condition and trial number, as well as the two respective main effects improved the general fit of the model to the data by comparing models with and without these effects (Dobson & Barnett, 2018). We included the variable of subject as a random effect. To analyze age effects in the child sample, we conducted an ANOVA including the interaction of condition and age in days, to see if they make more condition-specific choices with increasing age. Here we used the subsample of n = 67 children for which we knew the exact age.

To analyze children's answers to the question at the end of the experiment, we correlated the overall performance for each child with a question score. We assigned two points when children spontaneously answered the question "What do you think is in this hand?" correctly (correct animal in the type condition; length of the sticker strip in the size condition). We also counted as correct if they pointed to the correct example item. If children did not answer the question or provided a generic answer like "I don't know", we scored one point if the second forced-choice question was answered correctly. If both questions were answered incorrectly, no points were given. The correlation between the question score (ranging from 0 to 2) and the summed performance score (ranging from 0 to 6 correct trials) was performed using a Pearson moment correlation coefficient while controlling for age in years (pcor.test function in R; Kim, 2015).

Model Fit. We compared the a priori model predictions for the model, including Level 2 abstraction, the lesioned model only capable of Level 1 abstraction and a chance prediction (choice for each container at 50 %) to the children's and capuchins' choices for the hidden sample from the container with the small, high-value example item. Therefore, we computed the log-likelihoods for the test trials from each model. We used the differences in the Akaike Information Criterion (AIC) scores to compare the model predictions. $\Delta AIC > 2$ is generally considered strong support for the higher scoring model. For both species, AIC scores were first determined separately for the type and the size condition before the summed scores for each model were compared. We compared the overall performance in each condition to the model predictions for a single trial (with three evidence boxes). This assumes that children and monkeys treated every trial as independent and did not accumulate evidence and overhypotheses over trials.

2.4. Results

2.4.1. Reward preference testing

In the type comparisons, both species significantly preferred high-value items over equally sized low-value alternatives (Monkeys: $M = 0.86 \pm 0.12$, $t(16) = 11.78$, $p < 0.001$; Children: ($M = 0.94 \pm 0.18$, $t(79) = 22.33$, $p < 0.001$). The monkeys further preferred the least liked high-value item over equally sized pieces of medium-valued foods ($M = 0.90 \pm 0.06$, $t(16) = 25.35$, $p < 0.001$).[2] Both groups also significantly preferred large over small items (Monkeys: $M = 0.83 \pm 0.06$, $t(16) = 24.07$, $p < 0.001$; Children: $M = 0.83 \pm 0.32$, $t(79) = 9.11$, $p < 0.001$). In the mixed comparisons both groups expressed a significant preference for the small, high-valued items over the big, low-value option (Monkeys: $M = 0.97 \pm 0.04$, $t(16) = 53.46$, $p < 0.001$; Children: $M = 0.93 \pm 0.17$, $t(57) = 18.88$, $p < 0.001$; see SI for details).

2.4.2. Main experiment

Capuchin Monkeys. Monkeys were equally likely to choose the hidden sample from the container with the small high-value example in both conditions (paired $t(10) = 0.27$, $p = 0.79$), with their choices being at chance level in the type ($M = 0.52 \pm 0.06$, $t(10) = 1.10$, $p = 0.30$) as well as size condition ($M = 0.51 \pm 0.08$, $t(10) = 0.46$, $p = 0.65$; see Fig. 3B). Unlike in the preference testing, 7/11 monkeys expressed a side bias (same side in over 80 % of trials) regarding either the side of the experimenter's hand or the side of the container. They did not generally reach more frequently to the side of the small, high-valued test sample ($M = 0.52$; see SI for detailed results).

Children. 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, only the choices in the type condition were significantly different (type: $M = 0.71 \pm 0.24$, $t(39) = 5.70$, $p < 0.001$; size: $M = 0.45 \pm 0.22$, $t(39) = -1.45$, $p = 0.15$). The GLMM revealed no significant interaction between condition and trial number ($\chi^2(1) = 1.32$, $p = 0.25$), showing that children's choices in different conditions did not diverge more over trials. When excluding the non-significant interaction, the GLMM confirmed the significant effect of condition ($\chi^2(1) = 23.70$, $p < 0.001$). As expected, the children were not, in general, more likely to choose the sample from the container with the small, high-valued example over trials ($\chi^2(1) = 0.30$, $p = 0.58$). We further examined whether the children in the different conditions would make more differential, condition-specific choices with increasing age. The ANOVA revealed a marginally significant interaction of condition and age in days ($F(1) = 3.32$; $p = 0.07$; Fig. 4A). As expected, children were not, in general, more likely to choose the item from the small, high-valued container with increasing age ($F(1) = 0.12$; $p = 0.73$). Also, this ANOVA confirmed the overall effect of condition ($F(1) = 15.87$; $p < 0.001$).

When asked to predict what is in the chosen experimenter's hand after the last trial, 52.6 % of the children in the type but only one child in the size condition (2.6 %) answered the question correctly (see SI for more detail). Of the children that did not answer the first

[2] Some of this work has been published in the conference proceedings of CogSci 2019 (Felsche, E., Stevens, P., Völter, C. J., Buchsbaum, D., & Seed, A.M. "Exploring the use of overhypotheses by children and capuchin monkeys." In *CogSci*, pp. 1731–1737. 2019). The earlier version contained a few numerical errors which were corrected in this version of the paper. Further, we improved the computational model for the Level 1 formation. None of the changes had an impact on the results or conclusions. This version is to be seen as the correct one.

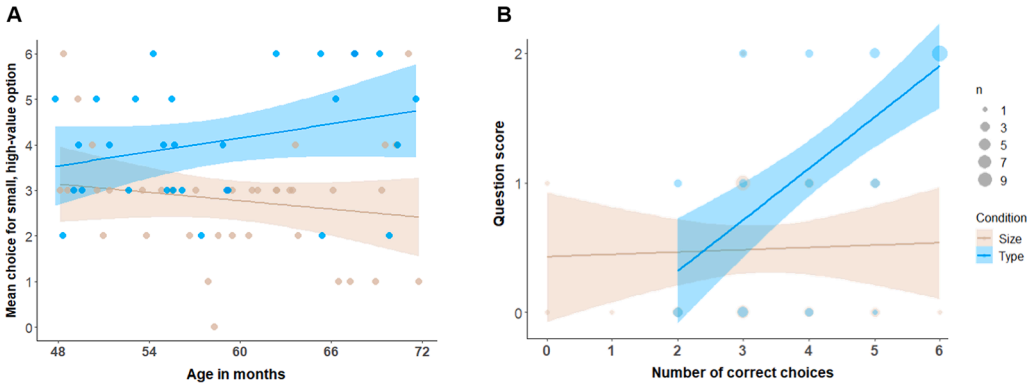

Fig. 4. A: Children's mean number of choices for the item from the container with the small, high-value example item in each condition depending on their age in months. B: Question Score for each participant given the number of correct choices during the main experiment (choice for container with small, high-value item in the type condition; and for container with large, low-valued item in the size condition). Regression lines are added to visualize correlations between the question score and the number of correct choices separately for the type and the size condition.

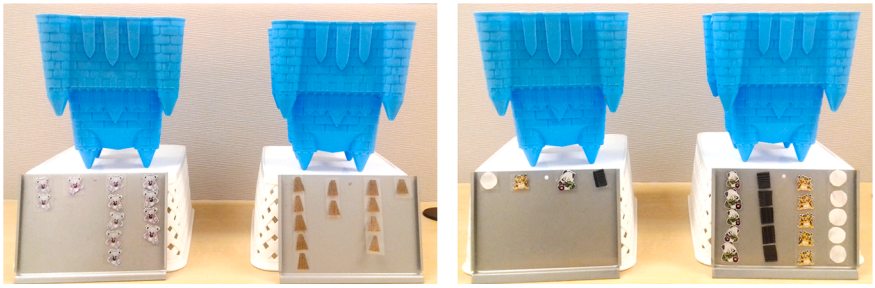

Fig. 5. Follow up evidence depicted with the materials used for children (left: type condition, right: size condition). Here, sticker strips of different lengths equipped with a small magnet at the back were sampled from the boxes onto metal frames. Four items were sampled from each container before the choice samples were retrieved, which the experimenter kept hidden in her hand (see Fig. 1).

question correctly about half gave the correct answer in the subsequent forced-choice question (size: 48.6 %, type: 53.3 %). There was a significant correlation of the question score and the overall performance when controlling for age ($r = 0.49$, $n = 74$, $p < 0.001$). Splitting up the data by condition, there was a significant correlation in the type ($r = 0.63$, $n = 38$, $p < 0.001$) but not in the size condition ($r = 0.04$, $n = 36$, $p = 0.81$, see Fig. 4B).

2.5. Model comparison

Capuchin monkeys' performance was best described by the random choice prediction, which fit their performance better than the Level 2 model ($\Delta AIC = 344.37$) and the Level 1 model ($\Delta AIC = 13.53$). When comparing both Bayesian models, the Level 1 model was superior to the Level 2 model in predicting the monkey data ($\Delta AIC = 330.84$). The children's results were best described by the predictions for a learner capable of Level 2 overhypothesis formation. The Level 2 model predictions fit the empirical data better than the Level 1 model predictions ($\Delta AIC = 7.12$) and the random choice ($\Delta AIC = 21.97$). Further, the Level 1 model predicted the data better than a chance model ($\Delta AIC = 14.85$).

2.6. Discussion

Overall, both the statistical results, as well as the comparison to the a priori predictions of a Bayesian model fit to children's preferences, indicate that children in our task were able to generalise at a second level of abstraction based on limited evidence. Children's choices in the type condition differed from their choices in the size condition, despite the evidence from the test containers being the same in both cases. This reveals the influence of the condition-unique evidence on their choices. Interestingly, children's performance did not significantly improve over trials. This effect broadly matched the model predictions, which only showed a slight change with more accumulating evidence. The marginally significant age effect suggests a possible improvement in abstract knowledge formation over the late preschool years. When comparing the children's performance in each condition to chance, they only made unambiguously overhypothesis-conforming choices in the type but not in the size condition. Likewise, when asked in the last trial, they could only verbally predict the identity of the chosen item in the type but not in the size condition. This condition imbalance was also to some extent present in the Level 2 model predictions, suggesting that the difference in the preference data is in theory sufficient to

explain this result. Nevertheless, the chance performance could indicate that children are either not able to form overhypotheses regarding the size or quantity of items, or have a strong prior assumption that, in general, items in the world are sorted by type (see General Discussion).

In contrast to children, the monkeys' chance level performance in both conditions suggests that they did not learn abstract rules regarding food distribution patterns across containers. This result is in line with previous negative results from experiments testing monkeys' abilities for abstraction and the notion that this ability is unique to humans (Penn et al., 2008) or great apes (Thompson & Oden, 2000). However, we decided to further investigate the reasons behind the monkeys' failure in this novel, training-free task design. Their failure on the second level of abstraction could be due to an inability to form abstractions about containers in general (Level 2 abstractions) or it could be rooted in a difficulty to infer the content distributions of each evidence container (Level 1 overhypotheses) based on the sampled evidence (even though in theory, both levels should be inferred simultaneously). Following the model comparison, the monkey's performance did not follow a learner capable of Level 1 abstraction but was purely random. However, given the proximity of the two predictions (50 % chance and ~ 59 % Level 1) and the possibility of other limiting factors such as sustained attention and required inhibition during the long sampling phase (Tecwyn et al., 2017), we conducted a second experiment to better understand the factors underlying the monkeys' choices.

3. Experiment 2: Abstraction within containers

Given the capuchin monkeys' failure to form Level 2 abstractions in the first experiment, we were interested in their ability to form Level 1 abstractions in a simplified task. Thus, we conducted a second experiment in which no generalization across containers, but only generalizations from samples to hidden food populations inside containers and vice versa was required. This simplified procedure moreover allowed us to test the generalization of the computational model to a second task and investigate the children's bias for type related rules further. Even though the ability for level 1 abstraction could be seen as a prerequisite to or at least a simpler form of the ability to form level 2 abstraction investigated in experiment 1, we chose the potentially counterintuitive order of experiments on purpose. Our aim in experiment 1 was to test a spontaneous ability for abstraction as shown by 9-month-old infants in Dewar and Xu (2010), which should extend the literature for non-human primates that is heavily based on training-intensive RMTS tasks. Further, research by Kroupin and Carey (2021) has shown that training on a first level of abstraction can have a strong influence on later performance at higher levels of abstraction in human children and adults (see also Smirnova et al., 2015; Obozova et al., 2015 for birds' positive results in RMTS after first level same/different matching to sample training).

In the second experiment, we presented monkeys and children with only two containers from which we sampled four evidence items, respectively. Now the choice items were sampled directly from these containers, so no generalization to new containers (Level 2) was required. However, participants forming abstractions at Level 1 would choose successfully according to the model predictions, while those failing to form any abstractions would choose at chance.

3.1. Method

3.1.1. Participants

Participants were 47 4- to 5-year-old children recruited at two local museums in Toronto ($M_{age} = 5.1$ years $\pm$ 0.7, two children, one 4-year-old and one 5-year-old, did not provide a birthdate, 25 female (53 %); type condition: n = 24, 13 female (54 %), $M_{age} = 5.0$ years $\pm$ 0.6, one 4-year-old with missing birthdate; size condition: n = 23, 12 female (52 %), $M_{age} = 5.2$ years $\pm$ 0.9, one 5-year-old with missing birthdate). Two additional children were excluded because they either ended the game early or due to experimenter error. The total sample of capuchin monkeys consisted of 13 individuals ($M_{age} = 7.38 \pm 4.7$, 4 females (31 %)). Ten had previously completed Experiment 1. Out of the 13 subjects, 9 completed both conditions, whereas two participated only in the size and two only in the type condition.

3.1.2. Procedure, design and analysis

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 (see Fig. 5). 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 items were drawn from both containers in a randomized order (two low and two high-value, for monkeys in 50 % of trials also two medium-, one high- and one low-value). Crucially, one container yielded only small (size 1) and the other only big (size 5) items. The hidden choice item was identical to one of the four types previously drawn from the container and of the same size as the other items from that container. In the type condition, items of a uniform type in sizes 1, 2, 4, and 5 were drawn from each container. Here, one container offered only low-valued and the other only high-valued items. The choice item was a randomly sized piece of the expected type for this container. Per condition, monkeys received 3 sessions of 8 trials each in a blocked design with condition order counterbalanced. The new sample of children received one session of 6 trials in a between-subjects design, beginning with a familiarization and a preference testing of two size and two type comparisons.

A second coder, blind to the purpose of the study, coded choices for 21.2 % of all sessions, randomly chosen. Interrater agreement was excellent for children (Cohen's kappa = 0.97, $p < 0.001$) and monkeys (Cohen's kappa = 0.94, $p < 0.001$). We measured the number of correct choices in each condition. In the type condition, this refers to the number of trials in which the participants chose the hidden item from the container with the high-valued example items. In the size condition, we counted the number of trials in which participants chose the item from the container with the large example items. We compared the results to chance for each condition and

species separately using one-sample *t*-tests in R. The children's preference testing was analyzed using one-sample *t*-tests. To analyze possible age and condition effects in the child sample we conducted an ANOVA with an interaction of condition and age in days using the aov function in R. To analyze trial effects we followed the GLMM procedure of Experiment 1 and used the emtrends function of the emmeans package (Lenth & Lenth, 2018) for post-hoc comparisons.

3.2. Computational model and predictions

We used the same computational model as in Experiment 1. However, as no abstraction across containers was needed, we only calculated Level 1 predictions. Like participants, the model received no evidence other than the four samples from each test container. All predictions were based solely on the types and sizes seen in the samples. Thus, the model considered only sizes 1 and 5 in the size condition but sizes 1, 2, 4, and 5 in the type condition.

Further, mainly low and high-valued items were included in the analysis, except for the monkeys' size condition, in which also medium items were presented during the evidence and thus included. For both children and monkeys the model capable of Level 1 abstraction tailored to the species' preferences predicted above chance performance in both conditions after only one set of evidence (see Fig. 6). Due to the stronger type compared to size preferences, the performance in the type condition was predicted to be slightly better than that in the size condition. Monkeys' predictions were more extreme than that of children due to their stronger inferred preferences.

3.3. Results, model comparison and Discussion

As in Experiment 1, the monkeys performed at chance level in both conditions (type: $M = 0.50 \pm 0.04$, $t(10) = 0.32$, $p = 0.76$; size: $M = 0.50 \pm 0.02$, $t(10) = 0.56$, $p = 0.59$) with only minimal individual variation due to the susceptibility for side biases.

In the preference testing, children significantly preferred animal over shape stickers ($M = 0.89 \pm 0.23$, $t(46) = 11.65$, $p < 0.001$) and larger strips of stickers over smaller ones ($M = 0.83 \pm 0.28$, $t(46) = 8.04$, $p < 0.001$). As these values are close to those obtained in the first study, we used the same inferred reward utilities for the model predictions.

In the main trials of Experiment 2, children performed significantly above chance in the type condition ($M = 0.74 \pm 0.21$, $t(23) = 5.41$, $p < 0.001$) but not in the size condition ($M = 0.55 \pm 0.20$, $t(22) = 1.19$, $p = 0.25$). A two-way ANOVA of condition by age in days showed that the interaction of age and condition was not significant ($F(1) = 1.19$; $p = 0.28$). Further, there was no significant effect of age across conditions ($F(1) = 2.77$; $p = 0.10$) but a significant effect of condition ($F(1) = 8.44$; $p = 0.006$) with better performance in the type compared to the size condition (see Fig. 6). Children's performance showed a significant trial by condition interaction ($\chi^2(1) = 5.59$, $p = 0.018$). The post-hoc analysis revealed that the slopes over trials significantly differed between conditions ($p = 0.02$). However, neither the improvement over trials in the type condition (slope = 0.21, 95 % CI [-0.02, 0.43]) nor the slightly negative slope in the size condition (slope = -0.16, 95 % CI [-0.36, 0.05]) were significant. As in the previous analyses the condition factor improved model fit ($\chi^2(1) = 8.84$, $p = 0.003$) but there was no main effect of trial ($\chi^2(1) = 0.06$, $p = 0.94$).

When comparing the fit of the Level 1 model and of the chance prediction, the monkeys' performance was far better described by a random choice between the sample items than by the model ($\Delta AIC = 454.02$). Interestingly, the same is true for the children's results, albeit with a smaller but meaningful advantage for the chance prediction ($\Delta AIC = 6.38$). This result originates from children's poor performance in the size condition in which they performed at chance level. When comparing the fit of the Level 1 model versus the chance predictions for the type condition only, a better fit by the Level 1 model predictions is revealed ($\Delta AIC = 20.99$). However, for the size condition the chance prediction fits the children's behavior better than the Level 1 model ($\Delta AIC = 27.37$). These results show that children's ability to optimize their behavior based on formed overhypothesis seems to be influenced by the dimension within which a general pattern can be detected. Whether this condition difference originates from an inability to form abstraction across some

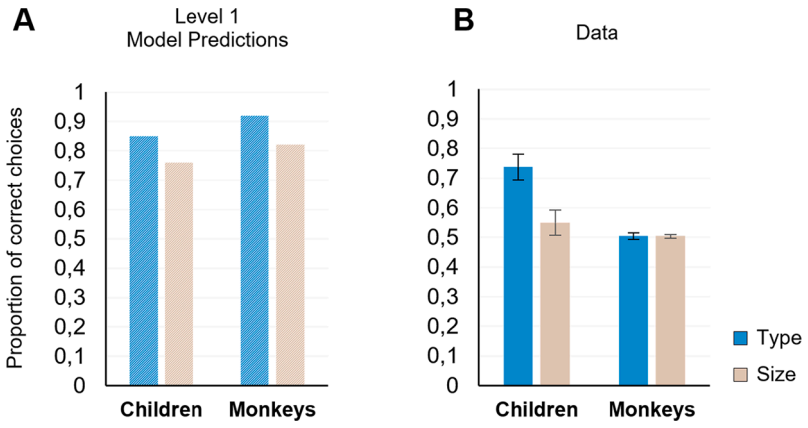

Fig. 6. Model predictions (A) and empirical results (B, M ± SE) for the correct choices of children and monkeys in the type (high-value item) and size (large item) condition.

dimensions or whether it results from motivational causes cannot be determined with the current evidence (see General Discussion). The capuchin monkeys in our study did not show evidence of spontaneous abstraction at any level. This is in line with previous results showing that monkeys only succeed after long training regimes even in same-different matching to sample tasks based on first-level abstraction (see General Discussion). Furthermore, neither species' data supports a strategy based on learning an association over trials between the boxes and less or more preferred items.

4. General Discussion

Over the course of two experiments, we investigated whether or not children and capuchin monkeys formed abstract rules given limited evidence by comparing their performance to the predictions of the Kemp et al. (2007) overhypothesis model. None of the capuchin monkeys showed the pattern predicted for a learner capable of forming abstract rules along the item size or type dimensions. In contrast, children in Experiment 1 treated the same test evidence differently when they had previously experienced that items were sorted by size or type. Their performance was well characterized by the predictions of a hierarchical Bayesian model capable of second level abstraction, fit to their species-specific reward preferences. This suggests that, in contrast to the capuchin monkeys, children formed overhypotheses at a second level of abstraction.

The monkeys tested with this paradigm did not show evidence of abstraction at the first level (from samples to hidden food populations in containers) nor the second level (across containers). These results are in line with low success rates achieved after long training regimes in previous studies on abstract concept formation and analogical reasoning in capuchin monkeys (e.g., Kennedy & Fragaszy, 2008; Thompson et al., 2016; Truppa et al., 2011; Wright et al., 2003). Our study extends previous findings with non-human animals to a novel, more ecologically valid task design based on abstract rules about the actual reward items presented in a sampling procedure. Related to the research on abstract patterns, our results also fit in with the picture found in studies using hierarchical stimuli. Here, previous research has shown that adult humans dominantly process stimuli on a global level (at least in some cultures; e.g., Navon, 1977; Oishi et al., 2014), whereas monkeys focus more on local features when confronted with hierarchical stimuli (e.g., a large square composed of small triangles; De Lillo et al., 2005; Fagot & Deruelle, 1997; Hopkins & Washburn, 2002). Thus, it is possible that during the evidence presentation, our monkeys failed to see the broader picture of, for example, "large – medium – small" but rather focused on the individual food items which naturally have a high salience to animals (see Fagot & Parron, 2010; Hochmann et al., 2017 for similar explanations of RMTS performance).

We can rule out some other possible explanations for monkeys' failure. Monkeys did not show a preference for the side exhibiting the small, high-value example item, which shows some understanding of the procedure as they were not simply trying to acquire the high-value test samples. Further, the sample was sufficient to detect significant food preferences, suggesting that the negative result was also not a sample size limitation and that the monkeys were, in principle, capable of making informed choices between food rewards hidden in the experimenter's hands. One could argue that the ABAB within-subject design in monkeys compared to the between-subject design in children potentially made abstract knowledge formation more difficult due to the switching between conditions. However, subjects' choices were not different from chance in either the first or the second block of either condition. Moreover, monkeys received more trials in each of the four blocks (12) than children in their single session (6). In contrast to children, monkeys only received three trials per session and experienced a gap of a few hours up to a few days between sessions. This break could perhaps have had a negative effect on their learning and evidence accumulation. However, both species' performance did not show any meaningful improvement over time, and the model comparison was not based on learning but assumed independent trials. While we manipulated the number of stickers on a strip in the size condition for children, monkeys saw a variation in the actual size of the food items. Both manipulations yielded clear results in the preference testing which shows that they were meaningful for the respective species. Thus it is unlikely that they have contributed to a species difference in performance. Nevertheless, there are many possible explanations for the monkey's chance level performance and their lack of fit to the predictions of overhypotheses-formation from a HBM. It is possible that the monkeys do not possess the ability to form abstract representations of same/uniform and different/mixed. Research on statistical inference (Eckert et al., 2018; Eckert et al., 2017; Rakoczy et al., 2014; Tecwyn et al., 2017) suggests that non-human primates are able to differentiate between various compositions of populations but it does not answer whether they also have an abstract representation of those compositions. Kroupin and Carey (2021) argue that even the gradual success of some non-human individuals in training-intensive RMTS tasks proofs they have the necessary abstract representational capacities, even if some authors promoting the lower-level perceptual account argue the contrary (e.g., Penn et al., 2008).

Still, it remains possible that other task demands inherent in the sampling procedure masked monkeys' abstract reasoning abilities. Capuchin monkeys and non-human great apes can infer the identity of a hidden item sampled from a visible population (Eckert et al., 2018; Eckert et al., 2017; Rakoczy et al., 2014; Tecwyn et al., 2017). However, despite performing above chance, the capuchin monkeys in Tecwyn et al., 2017 only chose correctly in 61 % of trials with uniform item populations (high vs low value) and 40 % of participants developed a side bias when choosing hidden samples from mixed populations. Tecwyn et al. (2017) suggested that additional cognitive demands, including object permanence, short-term memory and inhibitory control entailed in sampling paradigms, could contribute to the frequent appearance of side biases.

Future work could explore abstract reasoning with reduced task demands, e.g., allowing the participants to sample items themselves or touch the containers directly to indicate their choice. Further, one could explore the possibility of presenting monkeys (as the children here) only with two distinct reward categories of high and low-valued items, leaving out the medium category. Nevertheless, the novel approach taken here, in which participants do not need to be trained to make arbitrary judgments about abstract relations but simply need to secure the best rewards, is a promising avenue for future research. The findings from children provide evidence for abstract knowledge formation in 4- to 5-year-olds using Goodman's (1955) overhypothesis approach. The success following a small

amount of data with little improvement over trials was in line with the model predictions and further shows the speed of overhypothesis acquisition. The results further bridge a gap to the successful looking-time performance of infants in Dewar and Xu's (2010) study, as they show that older pre-schoolers master a highly similar choice paradigm on abstract rule formation from minimal sampled evidence. The marginally significant developmental improvement in our task between the ages of 4 and 5 is similar to that found in other studies using an RMTS paradigm (Christie & Gentner, 2014; Hochmann et al., 2017; Hoyos, Shao, & Gentner, 2016). However, it is unclear whether this potential developmental finding results from an improvement in abstraction abilities over time or older children are more able to handle other task demands (e.g., we found no age effect in the simplified second experiment). This overhypothesis-centered approach with limited verbal scaffolding and potentially reduced task demands could be extended to toddlers to bridge the gap across ontogeny.

Interestingly, children in both experiments performed above chance in the type condition but chose seemingly at random between the two hidden strips of stickers when presented with evidence according to a "containers are sorted by size" rule. The model predicts this imbalance between conditions to a certain extent, and the preference testing revealed slightly less extreme choices for the size compared to the type dimension. This is also emphasized by the mixed comparisons, in which a small, high-valued item is preferred over a larger but lower-valued item (as in monkeys). Thus, it could be that due to the task design or intrinsic preferences, children were less interested in or less focused on the size dimension, especially when the items simultaneously vary along both size and type dimensions. Thus, they may have focused on the fact that the containers in the size condition contained items that were a random mixture of types, which resulted in a random choice between containers. Further, it is possible that a strong prior bias for sorting items by their type over other dimensions contributed to the performance imbalance in the pre-schoolers. This bias could result from everyday experience that in general things in the world are mostly sorted by type and not size e.g., in grocery stores, on fruit trees or on toy shelves. Previous research based on hierarchical Bayesian models has shown how varying prior knowledge can influence the inferences children draw from experimental evidence (see Gopnik & Wellman, 2012 for a review). Given the variety of explanations for the condition difference, a closer investigation of children's ability for generalization along various object features in combination with manipulation of prior experience would be desirable. In addition, one could also vary the experimental protocol to look beyond fully uniform vs fully mixed distributions and introduce other distributions (e.g., 80:20, 60:40) to see if children's abstractions were absolute or more graded. Our studies' main aim was not to directly compare children and monkeys. Rather, we wanted to establish a novel and ecologically valid way of examining abstract reasoning performance that is suitable for participant groups that have previously failed the RMTS task. Nevertheless, the computational approach we applied in this study greatly improved the comparability of species and precisely formalized otherwise vague verbal predictions. In our study, we did not directly compare the species' performances due to the mostly inevitable differences in methodology (e.g., trial number; within & between-subject design, reward type). Instead, we compared each species' performance to a computational model that was tailored to each group's features (e.g., the number of reward value categories: high, medium, low vs high & low only) and the reward preferences established in previous preference testing. Only then did we descriptively compare the model fit of each species's performance. This way, species differences in methodology become less relevant for the comparison.

Despite many advantages, our implementation of the modeling approach to the current, novel context is limited by a lack of established guidelines. For example, while the modeling approach reduces subjectivity and experimenter bias, some parameters like the hypothesis space, the priors and the translation of the posterior predictions to behavior are chosen by the experimenter (Jones & Love, 2011). We used a choice rule previously established in child research (Lucas et al., 2014) and set the priors to be the most flexible and adaptive to the evidence. However, whether the choice rule also matches the monkey's decision-making process and whether a more biased prior towards uniform type distributions would have better captured the children's prior experiences remain questions for future research. As with most models, the HBM approach does not necessarily aim to represent the precise and exhaustive biological processes taking part in solving this task and lacks an explanation for how exactly new layers of abstraction are represented. Nevertheless, this approach has successfully described children's and adults' behavior in multiple instances (e.g. Griffiths et al., 2011; Lucas et al., 2014). This is thus a step forward in examining the influence of certain factors on the behavior of humans and potentially other species. Future studies could explore this approach further by comparing the Bayesian approach to connectionist and other computational models. Further, future endeavors should explore the reasons for failure to form overhypotheses in more detail. By assessing the prior assumptions of participants, and manipulating various aspects of the evidence presented, one could potentially differentiate whether e.g., a strong prior for certain overhypotheses rendered the evidence less effective or whether participants are better in forming overhypotheses over some dimensions than others.

In summary, we conducted the first direct test of the hierarchical Bayesian approach to overhypothesis formation described by Kemp et al., 2007 in children and non-human animals. We extended it to make choice predictions based on inferred item utilities. We have shown that this approach is a promising model for how children can form generalizations from sparse evidence. In addition, we extended the current literature on abstract knowledge formation in non-human animals to a novel, training-free, more ecologically valid task design in which food rewards are the subject of the targeted abstract patterns. Further application of extended computational models to empirical data of overhypothesis formation is desirable to understand its development over early childhood and to promote the understanding of possible species differences.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2022.101530.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
Benard, J., Stach, S., & Giurfa, M. (2006). Categorization of visual stimuli in the honeybee Apis mellifera. *Animal cognition, 9*(4), 257–270.
Brooks, D. I., Ng, K. H., Buss, E. W., Marshall, A. T., Freeman, J. H., & Wasserman, E. A. (2013). Categorization of photographic images by rats using shape-based image dimensions. *Journal of Experimental Psychology: Animal Behavior Processes, 39*(1), 85–92.
Carstensen, A., & Walker, C. (2017). The paradox of relational development is not universal: Abstract reasoning develops differently across cultures.
Christie, S., Gao, Y., & Ma, Q. (2020). Development of Analogical Reasoning: A Novel Perspective From Cross-Cultural Studies. *Child Development Perspectives, 14*(3), 164–170.
Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*(3), 356–373.
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.
Cook, R. G., Wright, A. A., & Drachman, E. E. (2013). Categorization of birds, mammals, and chimeras by pigeons. *Behavioural processes, 93*, 98–110.
Corballis, M. C. (2009). Do rats learn rules? *Animal Behaviour, 78*(4), e1–e2.
De Lillo, C., Spinozzi, G., Truppa, V., & Naylor, D. M. (2005). A comparative analysis of global and local processing of hierarchical visual stimuli in young children (Homo sapiens) and monkeys (Cebus apella). *Journal of Comparative Psychology, 119*(2), 155–165.
Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science, 21*(12), 1871–1877.
Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC Press.
Eckert, J., Call, J., Hermes, J., Herrmann, E., & Rakoczy, H. (2018). Intuitive statistical inferences in chimpanzees and humans follow Weber's law. *Cognition, 180*, 99–107.
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.
Fagot, J., & Deruelle, C. (1997). Processing of global and local visual information and hemispheric specialization in humans (Homo sapiens) and baboons (Papio papio). *Journal of Experimental Psychology: Human Perception and Performance, 23*(2), 429–442.
Fagot, J., & Parron, C. (2010). Relational matching in baboons (Papio papio) with reduced grouping requirements. *Journal of Experimental Psychology: Animal Behavior Processes, 36*(2), 184–193.
Flemming, T. M., & Kennedy, E. H. (2011). Chimpanzee (Pan troglodytes) relational matching: Playing by their own (analogical) rules. *Journal of Comparative Psychology, 125*(2), 207–215.
Garlick, D. (2010). *Intelligence and the brain: Solving the mystery of why people differ in IQ and how a child can be a genius*. Burbank, CA: Aesop Press.
Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review, 71*(2), 369–382.
Gentner, D. (2003). Why we're so smart. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5), 752–775.
Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child Development, 82*(4), 1173–1188.
Goddu, M. K., Lombrozo, T., & Gopnik, A. (2020). Transformations and transfer: Preschool children understand abstract relations and reason analogically in a causal task. *Child Development, 91*(6), 1898–1915.
Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.
Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. Proceedings of the 31st annual conference of the cognitive science society. 2759–2764.
Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin, 138*(6), 1085.
Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35*(8), 1407–1455.
Haun, D. B., & Call, J. (2009). Great apes' capacities to recognize relational similarity. *Cognition, 110*(2), 147–159.
Hespos, S., Gentner, D., Anderson, E., & Shivaram, A. (2021). The origins of same/different discrimination in human infants. *Current Opinion in Behavioral Sciences, 37*, 69–74.
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.
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.
Hopkins, W. D., & Washburn, D. A. (2002). Matching visual stimuli on the basis of global and local features by chimpanzees (Pan troglodytes) and rhesus monkeys (Macaca mulatta). *Animal cognition, 5*(1), 27–31.
Hoyos, C., Shao, R., Gentner, D. (2016). The paradox of relational development: Could language learning be (temporarily) harmful? In D. Grodner, D. Mirman, A. Papafragou, J. Trueswell, J. Novick, S. Arunachalam, S. Christie, C. Norris (Eds.), Proceedings of the 38th annual conference of the cognitive science society. Cognitive Science Society.
James, W. (1890), *Vol. 1. The principles of psychology*. New York: Henry Holt and Company.
Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and brain sciences, 34*(4), 169–231.

Katz, J. S., Wright, A. A., & Bodily, K. D. (2007). Issues in the comparative cognition of abstract-concept learning. *Comparative Cognition & Behavior Reviews, 2*, 79–92.
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), 167–175.
Kim, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods, 22*(6), 665–674.
Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797–2822.
Kroupin, I., & Carey, S. (2021). Population differences in performance on Relational Match to Sample (RMTS) sometimes reflect differences in inductive biases alone. *Current Opinion in Behavioral Sciences, 37*, 75–83.
Lazareva, O. F., Freiburger, K. L., & Wasserman, E. A. (2004). Pigeons concurrently categorize photographs at both basic and superordinate levels. *Psychonomic Bulletin & Review, 11*(6), 1111–1117.
Lenth, R., & Lenth, M. R. (2018). Package 'lsmeans'. *The American Statistician, 34*(4), 216–221.
Locke, J. (1847). *An essay concerning human understanding*. Kay & Troutman.
Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition, 131*(2), 284–299.
Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34*(1), 113–147.
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.
Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). *Rule learning by seven-month-old infants. science, 283*(5398), 77–80.
Mayer, C., Call, J., Albiach-Serrano, A., Visalberghi, E., Sabbatini, G., & Seed, A. (2014). Abstract knowledge in the broken-string problem: Evidence from non-human primates and pre-schoolers. *PLoS One, 9*(10), e108597.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology, 9*(3), 353–383.
Neiworth, J. J., London, J. M., Flynn, M. J., Rupert, D. D., Alldritt, O., & Hyde, C. (2017). Artificial grammar learning in tamarins (Saguinus oedipus) in varying stimulus contexts. *Journal of Comparative Psychology, 131*(2), 128–138.
Obozova, T., Smirnova, A., Zorina, Z., & Wasserman, E. (2015). Analogical reasoning in amazons. *Animal cognition, 18*(6), 1363–1371.
Oishi, S., Jaswal, V. K., Lillard, A. S., Mizokawa, A., Hitokoto, H., & Tsutsui, Y. (2014). Cultural variations in global versus local processing: A developmental perspective. *Developmental psychology, 50*(12), 2654–2665.
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and non-human minds. *Behavioral and Brain Sciences, 31*(2), 109–130.
Povinelli, D. J., & Penn, D. C. (2011). Through a floppy tool darkly Toward a conceptual overthrow of animal alchemy. In T. T. McCormack, C. Hoerl, & S. Butterfill (Eds.), *Tool use and causal cognition* (pp. 69–88). Oxford: Oxford University Press.
Premack, D. (2010). Why humans are unique: Three theories. *Perspectives on Psychological Science, 5*(1), 22–32.
R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical.
Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition, 131*(1), 60–68.
Ravignani, A., Filippi, P., & Tecumseh Fitch, W. (2019). Perceptual tuning influences rule generalization: Testing humans with monkey-tailored stimuli. *i-Perception, 10*(2), 2041669519846135.
Ravignani, A., Westphal-Fitch, G., Aust, U., Schlumpp, M. M., & Fitch, W. T. (2015). More than one way to see it: Individual heuristics in avian visual computation. *Cognition, 143*, 13–24.
Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of experimental child psychology, 94*(3), 249–273.
Seed, A., Hanus, D., & Call, J. (2011). Causal knowledge in corvids, primates, and children. *Tool use and causal cognition*, 89–110.
Seed, A. M., Call, J., Emery, N. J., & Clayton, N. S. (2009). Chimpanzees solve the trap problem when the confound of tool-use is removed. *Journal of Experimental Psychology: Animal Behavior Processes, 35*(1), 23–34.
Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*(4), 325–345.
Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology, 25*(2), 256–260.
Sonnweber, R., Ravignani, A., & Fitch, W. T. (2015). Non-adjacent visual dependency learning in chimpanzees. *Animal cognition, 18*(3), 733–745.
Spierings, M. J., & Ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences, 113*(27), E3977.
Tanaka, M. (2001). Discrimination and categorization of photographs of natural objects by chimpanzees (Pan troglodytes). *Animal cognition, 4*(3–4), 201–211.
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., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629–640.
Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences, 10*(7), 309–318.
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.
Thibaut, J.-P., French, R., & Vezneva, M. (2008). Analogy-making in children: the importance of processing constraints. Proceedings of the Annual Meeting of the Cognitive Science Society.
Thompson, R. K., Flemming, T. M., & Hagmann, C. E. (2016). Can old-world and new-world monkeys judge spatial above/below relations to be the same or different? Some of them, but not all of them. *Behavioural processes, 123*, 74–83.
Thompson, R. K., & Oden, D. L. (2000). Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. *Cognitive Science, 24*(3), 363–396.
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.
T. D. Ullman, C. L. Baker, O. Macindoe, O. Evans, N. D. Goodman, J. B. Tenenbaum, Help or hinder: Bayesian models of social goal inference, in: 22nd International Conference on Neural Information Processing Systems, 2009, pp. 1874–1882.
Vonk, J. (2015). Corvid cognition: Something to crow about? *Current Biology, 25*(2), R69–R71.
Vonk, J., Jett, S. E., Mosteller, K. W., & Galvan, M. (2013). Natural category discrimination in chimpanzees (Pan troglodytes) at three levels of abstraction. *Learning & behavior, 41*(3), 271–284.
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. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science, 25*(1), 161–169.
Walker, C. M., & Gopnik, A. (2017). Discriminating relational and perceptual judgments: Evidence from human toddlers. *Cognition, 166*, 23–27.
Wasserman, E. A., Castro, L., & Fagot, J. (2017). Relational thinking in animals and humans: From percepts to concepts. In J. Call (Ed.),. (Editor-in-Chief) American Psychological Association Handbook of Comparative Cognition (Vol. 2, pp. 359–384). Washington, DC: American Psychological Association.
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), 3–22.
Wright, A. A., Rivera, J. J., Katz, J. S., & Bachevalier, J. (2003). Abstract-concept learning and list-memory processing by capuchin and rhesus monkeys. *Journal of Experimental Psychology: Animal Behavior Processes, 29*(3), 184–198.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental science, 10*(3), 288–297.
Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological review, 114*(2), 245.
Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science, 26*(8), 1316–1324.