Cognition 245 (2024) 105721

Contents lists available at ScienceDirect

Cognition

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

How can I find what I want? Can children, chimpanzees and capuchin monkeys form abstract representations to guide their behavior in a sampling task?

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

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

ARTICLE INFO

Keywords:
Overhypotheses
Abstraction
Generalization
Animal cognition
Computational modeling
Cognitive development

ABSTRACT

Abstract concepts are a powerful tool for making wide-ranging predictions in new situations based on little experience. Whereas looking-time studies suggest an early emergence of this ability in human infancy, other paradigms like the relational match to sample task often fail to detect abstract concepts until late preschool years. Similarly, non-human animals show difficulties and often succeed only after long training regimes. Given the considerable influence of slight task modifications, the conclusiveness of these findings for the development and phylogenetic distribution of abstract reasoning is debated. Here, we tested the abilities of 3 to 5-year-old children, chimpanzees, and capuchin monkeys in a unified and more ecologically valid task design based on the concept of "overhypotheses" (Goodman, 1955). Participants sampled high- and low-valued items from containers that either each offered items of uniform value or a mix of high- and low-valued items. In a test situation, participants should switch away earlier from a container offering low-valued items when they learned that, in general, items within a container are of the same type, but should stay longer if they formed the overhypothesis that containers bear a mix of types. We compared each species' performance to the predictions of a probabilistic hierarchical Bayesian model forming overhypotheses at a first and second level of abstraction, adapted to each species' reward preferences. Children and, to a more limited extent, chimpanzees demonstrated their sensitivity to abstract patterns in the evidence. In contrast, capuchin monkeys did not exhibit conclusive evidence for the ability of abstract knowledge formation

1. Introduction

Humans greatly benefit from their ability to detect commonalities between two or more objects, relations, processes, or situations, allowing them to extract general patterns that go beyond the immediate sensory input. This ability for abstraction enables humans to adaptively transfer knowledge and problem solutions from one situation to another without the need to remember situation-specific details. For example, imagine that upon acquiring a large collection of vinyl records, you embark on a quest to curate '60s music for an upcoming event. After examining the first five records in one box, all dated 1982, and discovering only records from 1969 in another box and some from 1975 in a third, you recognize an abstract pattern - within each box, all records are from the same year. This realization about how the records are sorted in boxes enables you to optimize your search strategy: a brief inspection of the first record in each box offers a strong indication of whether it contains records from the '60s, eliminating the need to look at every individual item. This kind of abstract reasoning is central to human intelligence, plays an essential role in the evolution of human culture, and is crucial for a variety of human-unique accomplishments in

* Corresponding author at: Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.
E-mail addresses: elisa_felsche@eva.mpg.de (E. Felsche), christoph_voelter@eva.mpg.de (C.J. Völter), esther.herrmann@port.ac.uk (E. Herrmann), ams18@st-andrews.ac.uk (A.M. Seed), daphna_buchsbaum@brown.edu (D. Buchsbaum).
[1] Equal contribution

https://doi.org/10.1016/j.cognition.2024.105721
Received 9 July 2023; Received in revised form 23 December 2023; Accepted 10 January 2024
Available online 22 January 2024

social interaction, language, mathematics, arts, and teaching (Brand, Mesoudi, & Smaldino, 2021; Gentner, 2003; Tomasello, 2020; Wasserman & Young, 2010). Some believe it is a relatively late-developing skill, dependent on language or symbols and only present in humans (e.g. Gentner, 1988, 2003; Holyoak & Lu, 2021; Penn, Holyoak, & Povinelli, 2008; Tomasello, 2020). Others argue that the ability for abstraction as a powerful and efficient learning tool is relevant for fast knowledge acquisition in infancy (e.g. Kemp, Perfors, & Tenenbaum, 2007; Xu & Tenenbaum, 2007b; Yin & Csibra, 2015). Moreover, the ability to extract recurring patterns in different contexts and transfer adaptive behaviours across situations could be advantageous for many species in light of evolutionary pressures to find food and reproduce, which unite disparate perceptual features according to their function or utility.

1.1. The abstract concepts of sameness and difference in animals and children

As outlined above, detecting similarities and differences between entities is crucial for forming abstract concepts. Thus, the abstract concepts of *same* and *different* itself play a central role in human cognition. However, their developmental timeline and the extent to which abstract reasoning is uniquely human continue to be debated. Human infants and a variety of other animals, including insects (Giurfa, Zhang, Jenett, Menzel, & Srinivasan, 2001), can learn to expect a specific event or engage in a particular action when stimuli are perceived to be the same or can learn to choose matching stimuli (e.g. Ferry, Hespos, & Gentner, 2015; Hochmann, Carey, & Mehler, 2018; Hochmann, Mody, & Carey, 2016; Wasserman, Castro, & Fagot, 2017). However, success in these tasks, including the transfer of the training performance to a test phase with novel stimuli, may not require flexible, fully abstract representations but could stem from simpler perceptual matching processes (e.g. Kroupin & Carey, 2022a; Penn et al., 2008; Zentall, Andrews, & Case, 2018).

In the more difficult relational matching-to-sample task (RMTS), participants are required to match not individual stimuli but instead the relations (same/different) between stimuli in pairs or arrays (XX matches AA but not BC. XY matches BC but not AA; Premack, 1983), which represents a second level of abstraction. Children under the age of 5 perform poorly on RMTS tasks (e.g. Hochmann et al., 2017). Their bias for conflicting object matches (e.g., choosing XX matches XB over AA) suggests that children only start attending to common relational structures in late preschool or early school age (Gentner, 1988; Rattermann & Gentner, 1998).

However, slight task modifications like presenting multiple samples (e.g. do XX and YY match XY or AA; Christie & Gentner, 2010), labels for samples (e.g. “this is a truffet, which one is also a truffet?”, Christie & Gentner, 2014), larger item arrays (Hochmann et al., 2017), causal framing (Goddu, Lombrozo, & Gopnik, 2020), or design cues drawing attention to relations between the stimuli (Walker, Rett, & Bonawitz, 2020) enable 3- and 4-year-olds to succeed in relational matching tasks. After learning that either only “same” or only “different” object pairs activate a blicket detector toy, English-speaking 18-month-olds outperformed 3-year-olds when choosing between new objects to put on the toy (Walker, Bridgers, & Gopnik, 2016). Mandarin-speaking children, however, succeed in choosing relational matches at both ages (Carstensen et al., 2019). These findings suggest that abstract relational reasoning may develop early but that learned, culture-dependent biases may shift children’s focus to individual objects. Thus, learned biases, rather than a lack of capacity for abstraction, may cause failures in RMTS tasks in 3- to 5-year-olds (Hoyos, Shao, & Gentner, 2016; Kroupin & Carey, 2022b; Walker et al., 2016). Supporting this argument, Kroupin and Carey (2022a, 2022b) found that a brief match-to-sample (MTS) training based on less attended object dimensions (like number or size) improved RMTS scores in 4-year-olds and adults, while an MTS training based on the objects’ shape and/or color did not. The RMTS task also represents a challenge for non-human animals (henceforth animals). Only a few monkey species have shown success on the task (e.g., capuchin monkeys (Truppa, Piano Mortari, Garofoli, Privitera, & Visalberghi, 2011), baboons (e.g. Fagot & Thompson, 2011)) after extensive training or the presentation of larger item arrays while still showing a drop in performance when new stimulus sets are introduced (Wasserman et al., 2017). Thus, animals seem to rely primarily on slow perceptual learning processes based on the specific stimulus combinations or on comparing their perceptual variability of the stimulus arrays (Penn et al., 2008; Wasserman et al., 2017). In birds and great apes, prior MTS experience and language or symbol training lead to faster success in RMTS tasks (Obozova, Smirnova, Zorina, & Wasserman, 2015; Premack, 1983; Smirnova, Zorina, Obozova, & Wasserman, 2015; Thompson, Oden, & Boysen, 1997) that is partly robust against conflicting perceptual matches (Vonk, 2003). This suggests that despite the apparent species differences (e.g. between monkeys and apes), having acquired a relevant symbol system (Gentner, Shao, Simms, & Hespos, 2021; Premack, 1983) or at least extensive prior exposure to the relation of sameness (Smirnova et al., 2015) supports the representation of abstract relations. However, even for those results, lower-level explanations have not been entirely ruled out (e.g. Dymond & Stewart, 2016; Penn et al., 2008; Vonk, 2015).

Similar to the argument for children, animals’ poor RMTS performance may not indicate a lack of abstract reasoning capacity. Inductive biases to pay attention to and assume meaning of other stimulus features, such as salient or previously relevant object features like shape, color or location, could influence animals’ relational responses. Numerous repetitions or specific training might shape their behavior to match the experimenter’s expectations (Carstensen & Frank, 2021; Kroupin & Carey, 2021).

Despite its popularity, the RMTS task presents an arbitrary scenario, especially for animals. The procedure often involves geometric shapes with little meaning shown on computer screens. Christie (2021) argues that some interest in the individual stimuli is necessary to detect relations between them. Human RMTS performance improves when meaningful stimuli (words vs. random letter strings) are presented, supporting the importance of stimulus choice for abstraction (Flemming, Beran, Thompson, Kleider, & Washburn, 2008). The RMTS as a binary forced-choice task, where a decision for one option makes another inaccessible, lacks ecological plausibility. Moreover, any learned response strategy here has to overcome the non-human primates’ strong focus on spatial solution strategies (e.g. Flemming & Kennedy, 2011; Haun, Call, Janzen, & Levinson, 2006) as apparent in the occurrence of side biases (e.g. Flemming, 2006). Given the RMTS task’s low-ecological validity, its susceptibility to lower-level explanations, and the ambiguity in its interpretation, we turned to another perspective on abstract reasoning that has more prominently been examined in developmental psychology. Further, we use computational modeling to achieve a more informative cross-species comparison.

1.2. Overhypotheses and Hierarchical Bayesian Models

Strongly related to the traditional formalization of “same-different” concepts is the notion of overhypotheses (Goodman, 1955) which offers a different perspective on testing abstract concepts and is a well-established term in computational, cognitive, and developmental psychology. Similar to the vinyl record example from above, Goodman (1955) illustrated this concept with a thought experiment on bags filled with marbles. Out of the first bag, you blindly draw some marbles, which turn out to be all red. From the second bag, you sample only blue, and from the third, only black marbles. With each draw, you become more confident in forming a first-order generalization about each bag’s marble distribution, for example, “the third bag contains only black marbles”. Further, you can extract the commonality of the events and form a second-order generalization, the overhypothesis: “Within a bag, all marbles have the *same* color”. This overhypothesis allows powerful predictions about the color of all marbles in a new bag after sampling

just one (Goodman, 1955). Similar to the bags example, one can imagine the marble-filled bags as trees growing fruit. If, after visiting a few trees, an animal can form the overhypothesis that trees generally carry a uniform fruit type, only one bite of a fruit from a novel tree is sufficient to determine whether to spend more time and energy foraging in this tree. In contrast, after lifting stones on the ground, an animal may learn the overhypothesis that a variety of insects can be hiding under any given rock. In this case, the first insect you see under a new stone does not make you sure about the type of the next insect you will find in this location. Similarly, learning that bags are each filled with a mix of colours means that the color of the next marble from a specific container is less predictable.

Conceptualizing abstract generalizations as overhypotheses represents them as theories (or hypotheses) that shape and constrain the hypothesis space at more specific levels (Kemp et al., 2007). For example, given that trees grow a uniform fruit type (overhypothesis), it is likely that this specific tree bears only apples but unlikely that it grows a mix of apples and cherries (lower-level hypothesis). While an RMTS task tests the possession of abstract knowledge, assuming concepts "same" and "different" are already in place and readily applied to displayed object pairs, the overhypothesis framework quantifies the genesis of abstractions from evidence and their application to novel instances. By so doing, it can explicitly investigate how features of the evidence at the concrete level might impact or facilitate such inferences - such as the contrasting impact of recently experiencing feature-based or relation-based rules on performance in the RMTS task in children described above (Kroupin & Carey, 2021). This is a desirable feature for a theoretical framework exploring how abstractions can be derived and used.

Probabilistic hierarchical Bayesian models (HBM; e.g., Kemp et al., 2007; Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths, & Kemp, 2006) show how, in theory, this structured abstract knowledge could be acquired after observing limited data, due to simultaneous Bayesian inference at multiple levels of abstraction. These overhypotheses are updated in light of new evidence across a broad range of situations (e.g., multiple trees and kinds of trees) and thus can be learned faster than hypotheses concerning more specific situations (this particular tree; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). The HBM also incorporates how overhypotheses determine the probability distribution for the hypothesis space at lower levels. For example, knowing that items within bags are highly uniform, one assigns a high probability of sampling another item of the previous type but a low probability of sampling any other type. However, if one samples from a bag and entertains the overhypothesis that bags usually have a mix of items, one would assign similar (low to moderate) probabilities to multiple types of items to be the next sample. Thus, as a computational formalism, they capture the ability for rapid inferences and wide-ranging predictions when encountering new but related situations.

At the extreme end of the continuum of uniformity, these overhypotheses map well onto the concepts of 'same' and 'different'. However, it should be noted that they remain inherently probabilistic in nature. While relations (e.g. larger than, same, middle of) *can* be represented as overhypotheses, there are other ways in which relations could be conceptualised – for example, as all-or-nothing propositional rules. Our study will examine how participants might form abstractions concerning the uniformity of different populations from evidence but leaves open whether the nature of the representations generated can be termed relational as such.

This modeling approach has successfully characterized human behavior in different inductive learning scenarios (Gopnik & Wellman, 2012) like language acquisition (Xu & Tenenbaum, 2007a, 2007b), inferences about social groups (Kemp et al., 2007) or causal learning (Lucas & Griffiths, 2010). For example, Lucas, Bridgers, Griffiths, and Gopnik (2014) showed that consistent with hierarchical models with different a priori overhypotheses, preschoolers flexibly learned that a conjunction of two objects was causally necessary to activate a machine. Meanwhile, older children and adults made inferences consistent with having previously formed an overhypothesis that individual objects hold causal power. Similarly, an acquired object focus that harms RMTS performance could be represented as a prior overhypothesis.

In studies by Sim and Xu (2015, 2017), 2- and 3-year-old children and 17- to 20-month-old toddlers formed a second-order generalization regarding the functionality of objects. They were presented with three sets, each including two machines and one activator block that matched the machines in either color or shape (depending on the condition). In the subsequent test, both age groups chose a correct novel activator block (color or shape match) for familiar (first-order) and novel machines (second-order generalization). Unlike the older children (Sim & Xu, 2017), toddlers needed guidance from an experimenter or parent to generate the required evidence for later generalizations but failed to do so in independent free play (Sim & Xu, 2015).

Utilizing the idea of overhypotheses and testing even younger infants, Dewar and Xu (2010) found positive evidence for abstract concepts in 9-month-old infants. In an evidence phase, the experimenter sampled four objects from each of three boxes. The items within a box had the same unique shape but varied in color (e.g. 4 spheres, 4 cubes, 4 stars). Then, a new box was presented from which the experimenter sampled two items of the same shape or two items of differing shapes. Infants looked longer at the latter sample, suggesting that they noticed the apparent violation of the previously learned overhypothesis: "objects within a box have the same shape". This suggests that already preverbal infants can form abstract concepts based on limited sampled evidence. However, whether this ability can support decision-making in a choice situation and translates to the later preschool age range is unclear.

Inspired by this paradigm and the original overhypothesis thought experiment (Goodman, 1955), Felsche, Stevens, Völter, Buchsbaum, and Seed (2023) conducted a choice study with 4- and 5-year-old children and capuchin monkeys. They compared the empirical performance of each species to an HBM (adapted from Kemp et al., 2007), capturing the choices participants should normatively make if they had learned the relevant overhypotheses. Participants saw sampled evidence indicating either that containers hold items of uniform type (e.g. A: banana, B: carrot, C: apple) but varying size, or that items are sorted by size (e.g. A: small items, B: large items, C: medium size), but that each container offers a mix of item types. Subsequently, participants of both conditions were presented with two new test boxes: from container D, the experimenter sampled a small, high-valued item, and from E, a large but low-value item. Next, participants could choose between a new hidden sample from each container. The HBM predicted that if participants inferred the overhypothesis that items are sorted by type, they should choose the sample from D to obtain another high-valued item (of a random size). In contrast, they should select the item from E to secure another large reward (of a random type) when they have seen that items are sorted by size. Children showed the expected difference between conditions, and their performance was well predicted by the HBM capable of overhypothesis formation. However, the capuchin monkeys showed no evidence of overhypothesis formation. While these findings could indicate that capuchin monkeys lack a capacity for abstract concept formation, the passive sampling procedure likely imposed additional task demands regarding abilities for inhibition, object permanence, and working memory that might have especially impacted the capuchin monkeys' performance (Tecwyn, Denison, Messer, & Buchsbaum, 2017).

1.3. The current study

To accurately investigate the abilities of non-human primates and young children to engage in abstract reasoning, we need to use a less demanding test environment with a more naturalistic choice situation in which the evidence-gathering process is self-determined by the subject. In the current study, we apply the idea of overhypothesis and the HBM approach to the comparative study of abstract relational reasoning. In

contrast to the classic RMTS design, we implement a more ecologically valid approach with reduced task demands based on a participant-led and self-conducted sampling procedure. Rather than using a binary forced-choice procedure, we introduced a natural foraging scenario in which we measured the efficiency of a search through food patches with actual reward items of differing values. Here, participants did not need to explicitly detect the commonality of "sameness" or "difference" in arbitrarily displayed item pairs but could instead acquire an overhypothesis about the commonalities of the container contents (items within containers are of the *same* type or *different* types) over time when sampling their own evidence. To further reduce task demands, we varied only the distribution of item types across containers instead of contrasting the variation in two item dimensions (as in Dewar & Xu, 2010; Felsche et al., 2023). In statistical reasoning paradigms, infants (e.g. Denison & Xu, 2014; Gweon, Tenenbaum, & Schulz, 2010; Téglás et al., 2011), preschoolers (Denison, Bonawitz, Gopnik, & Griffiths, 2013; Girotto, Fontanari, Gonzalez, Vallortigara, & Blaye, 2016) as well as non-human primates (Eckert, Call, Hermes, Herrmann, & Rakoczy, 2018; Eckert, Rakoczy, & Call, 2017; Rakoczy et al., 2014; Tecwyn et al., 2017) have shown sensitivity to the composition of item populations within containers and the resulting probability for a sampled item to be of a specific item type. However, none of these studies have investigated the generalization of item distribution patterns across containers.

In the current study, we included 3-, 4- and 5-year-old English-speaking children as this age range usually marks the transition of failure in the classic RMTS task at age 3 to mostly successful performance at five years of age (Christie & Gentner, 2010; Hochmann et al., 2017; Walker et al., 2016). However, as outlined above, slight task modifications have shown success in 3-year-olds and thus suggest that abilities for abstraction are present at that age. The current study explores further the conditions under which preschool children show spontaneous abstract concept formation. Additionally, we tested symbol- and language-naïve chimpanzees (*Pan troglodytes*) and capuchin monkeys (*Sapajus apella*) in our study. Members of both species have not only reached above-chance level performance in RMTS tasks (e.g. Flemming et al., 2008; Premack, 1983; Thompson et al., 1997; Truppa et al., 2011) but also showed some intuition for abstract patterns in relational reasoning tasks based on spatial or size relations (Flemming & Kennedy, 2011; Haun & Call, 2009; Kennedy & Fragaszy, 2008). In these tasks, subjects had to choose the cup from a set of three with the same relative but not necessarily absolute size or position as a visibly baited cup in the experimenter's set (e.g. largest cup). In all experiments on abstract relational reasoning, chimpanzees typically succeeded at higher rates and within fewer trials than capuchin monkeys. Thompson and Oden (2000) even proposed that the line differentiating species capable of abstract reasoning from those solely relying on first-level cues should not be drawn between humans and other primates but between apes and monkeys. However, the sample sizes in these studies usually involve less than ten individuals per species, and unlike the capuchin monkeys, most chimpanzees tested received prior language or symbol training. Including non-enculturated chimpanzees and capuchin monkeys in the current study will provide crucial evidence on non-human primates' abilities for abstraction and give insights into whether their difficulties in RMTS might reflect context and bias rather than ability (Kroupin & Carey, 2021).

In the current study, we provided chimpanzees, capuchin monkeys, and 3- to 5-year-old children with the opportunity to sample their own evidence that either suggested that containers are filled with reward items of a uniform type (all items in a container are of high or all of low-value, like the fruit trees or uniformly coloured marble-bags) or that each container offers a balanced 50:50 mixture of high- and low-valued item types (like the insects under stones or the mixed-coloured marble bags). In a subsequent test situation, all participants were simultaneously presented with two new containers, which, unbeknownst to the participants, were both filled entirely with low-valued items, regardless of the condition. Suppose participants in the uniform condition learned the overhypothesis that containers provide items of the same type. In that case, receiving one low-valued item should, in theory, motivate the learner to consider switching away from this container, which likely contains only low-valued items, and explore the second container for potential high-valued items. In contrast, participants that previously experienced that each container offers the same mix of item types should on average be more persistent with the first container and not get discouraged to the same degree by the first few low-valued items. As in Felsche and colleagues (2023), we compared the participants' behavior to the predictions of a probabilistic hierarchical Bayesian model fitted to the species' item preferences and equipped with a choice rule for when to switch from one container to the other.

While abstract reasoning is often seen as a domain-general ability (Gentner, 2003; Penn et al., 2008), task performance nonetheless shows sensitivity to slight task modifications, as observed for variations of RMTS tasks. Further, for a given task the required response behavior may better match the behavioural repertoire of some species over others (e.g., foraging for items on the ground vs. pressing a button on a machine), introducing varying task demands outside of the cognitive ability in focus. To ensure the generalizability of the results and account for varying peripheral task demands across species, we presented three versions of the task: foraging for items hidden in material-filled buckets, lifting cups to uncover items, and operating a button to dispense items from a machine. We chose these presentations because each had pragmatic advantages and disadvantages. The machine version builds on previous findings of successful causal reasoning about puzzle boxes in both children and primates (e.g. Schulz, Kushnir, & Gopnik, 2007; Tennie et al., 2019) but is arguably the least ecologically valid for non-human species. The cups procedure clearly displayed the overall number of items available. The material-filled buckets were perhaps the most ecologically valid for the primates; however, the most challenging to determine which items had been sampled (see below).

The overall procedure and reward distributions were identical across versions. Only the presentation of the rewards and actions required to sample the rewards differed. The use of varied materials further facilitated a within-subject design for non-human primates, minimizing potential carry-over effects of learned overhypotheses from one session to the next.

2. Experiment 1

2.1. Method

2.1.1. Participants

Children. A total of 212 children between the ages of 3 and 5 was included in our final sample (106 female, M_{age} = 54.78 months ±10.36 SD; see SM, Table S1). The data collection took place at two local museums in Toronto (the Royal Ontario Museum and the Ontario Science Centre). An additional 16 children were excluded from the analysis due to experimenter error (6), interference by parents (4), apparatus error (3), their wish to stop early (1), emptying both test containers simultaneously (1) or switching without any evidence (1). For the separate preference testing, we collected data from an additional 40 3- to 5-year-old children (19 female, M_{age} = 54.83 months ±10.08 SD) tested at the same two museums. Three additional children were excluded from the analysis due to interference by family members (2) or misunderstanding of the task materials (1). The study was planned and conducted following ethical guidelines. It was approved by the School of Psychology and Neuroscience ethics committee 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.

Capuchin Monkeys. Overall, 22 capuchin monkeys (*Sapajus* sp.)

participated in this study (9 female, $M_{age} = 9.27$ years ±4.08 SD; see SM, Table S2). All but one are zoo-born and mother-raised. The capuchin monkeys are housed in two groups at the Living Links to Human Evolution Research Centre at Edinburgh Zoo. The animals have access to a large outdoor and indoor enclosure and are cohoused with squirrel monkeys (*Saimiri sciureus*) with whom they share their natural environment. The monkeys were never food or water restricted.

Chimpanzees. We collected data from 30 chimpanzees (17 female (57%), $M_{age} = 21.2$ years ±7.97 SD; see SM, Table S3) at the Sweetwaters Chimpanzee Sanctuary in the Ol Pejeta Conservancy in Laikipia county, Kenya. The chimpanzees live in two separate social groups and spend their day outside in large outdoor enclosures. They are provided with water ad-lib and fed three times daily. The experimental protocol and study design for the apes and monkeys was approved by the School of Psychology and Neuroscience ethics committee at the University of St Andrews, the local ethics committee at the chimpanzee sanctuary, Kenya Wildlife Service, and the Kenyan National Council for Science and Technology.

2.1.2. Materials

For all species, the experiment involved high- and low-value reward items that could be sampled from containers. For the chimpanzees, we used food items whose value was based on the caregiver's judgment (high: apple & banana, low: orange & carrot in evidence, raw sweet potato in test). Capuchin monkeys' rewards were based on a previous preference testing conducted in the same group (high: date & peanut, low: carrot & eggplant (evidence), zucchini (test); see Felsche et al., 2023 for details). For children, high-valued rewards were yellow and green balls that could subsequently be rolled down a marble run while producing an engaging sound. We used cubes that could not be inserted in the marble run game as low-valued items. A preference testing confirmed the relative value of the items (see SM for details).

We presented all participant groups with three versions of the task (see Fig. 1). Whereas children engaged with the materials on tables or on the floor, non-human primates sampled their rewards through a metal mesh or plexiglass barrier.

Machine version. Uniquely coloured and shaped machines released a reward upon pressing a button on the front of the machines.

Cup version. Participants could find rewards by knocking over cups attached to a board. For children, all 10 cups were randomly distributed on a rectangular board. For non-human primates, the 10 cups were arranged in a row to be accessible through the barrier.

Foraging Version. Participants could retrieve rewards from containers filled with other materials (children: packing peanuts; non-human primates: saw dust). While the setup for chimpanzees and children led them to sample rewards one by one, the capuchin monkeys often swept out most containers' contents with one arm movement. This impeded their opportunity to notice the items individually. Thus, we presented the monkeys with a second foraging version in which plastic barriers subdivided each container. We placed a reward covered by cut straw in each of the ten emerging compartments.

2.1.3. Procedure

Each session consisted of 4 evidence trials followed by the test situation. In each evidence trial, participants sampled items from a new container. Depending on the version, "container" represents a machine, board with cups, or bucket (see Materials). Each container held ten items. To prevent children from implicitly learning a game rule that they

Fig. 1. Photos of the test situation for each version (rows) and species (columns) as presented during the experiment. Left: children, Middle: capuchin monkeys, Right: chimpanzees; from top to bottom: machine version, cup version, foraging version (for capuchin monkeys, the second foraging version is depicted).

must empty a container before moving on, the experimenter changed evidence containers after a counterbalanced item number or time criterion (dependent on condition, see SM for details) that was not communicated to children. The non-human primates were allowed to empty all ten items in each evidence trial to avoid frustration caused by taking away food. We adapted the computational model to these species-specific amounts of evidence.

In a given session, participants were either presented with the uniform or the mixed condition (see Fig. 2). In the *uniform condition*, the sampled rewards were uniform or all the "same" within an evidence container but different across containers. In two evidence trials, the containers were filled with only high-valued items, and in the other two trials with only low-valued items. In the *mixed condition*, each evidence container offered an equal mix of five high- and five low-valued items (see Fig. 2). In both conditions, we counterbalanced the order of container and reward types across participants so that they experienced a low and a high-valued reward type in the first two evidence trials and another two reward types in the last two evidence trials. We also ensured a random sampling from each container (see SM for details).

After four evidence trials, participants moved on to the test situation in which two containers were presented simultaneously. The experimenter told the children that she had to do some other task but that there are two more containers left and that they should try to find more marbles to play the game. For the non-human primates, the experimenter set up both containers and ensured each participant had seen both before moving them simultaneously in reach of the participant. Each test container held ten low-valued items in all species, versions, and conditions. This ensured that participants would only find low-valued items, whichever container they started sampling from. After the participants switched containers, indicated the wish to leave the testing area or one minute without engagement with the test containers had passed, the session ended, and participants got rewarded with some high-valued items.

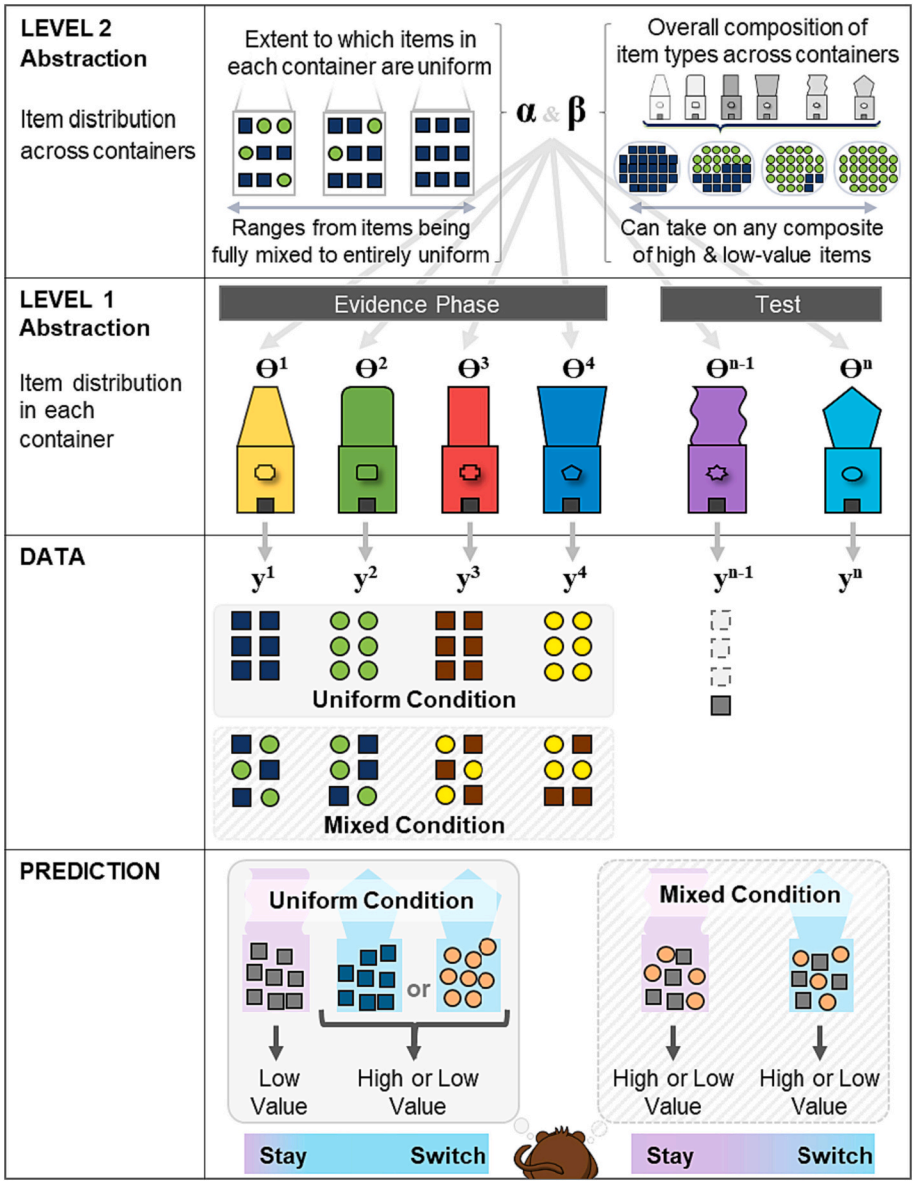

Fig. 2. Hierarchical Bayesian model of overhypothesis formation adapted for the current study. The parameters α and **β** describe an overhypothesis at the second level of abstraction: α represents the extent to which item types in containers, in general, tend to be uniform vs. mixed, and **β** captures the type variability across all containers. Type distributions of a specific container ($\boldsymbol{\theta}^i$, Level 1 abstraction) are constrained by overhypotheses at Level 2 and, in turn, constrain the items $\boldsymbol{y}^i$ sampled from that container. Squares represent low-valued items and circles represent high-valued items. In their choice to switch, a learner puts the expected values of the item distribution in each test container into relation.

2.1.4. Design

We applied a 2 (condition: mixed or uniform) x 3 (version: machine, cups, foraging) design in all species.

Children. We applied a between-subjects design where each child received one session. This session was run in one of the six condition x version combinations. Each of the six groups consisted of 33 to 37 participants (see Table S1).

Chimpanzees. Each chimpanzee received three sessions, one for each version. The first two sessions were run in the cup and foraging versions, and the third session was always run in the machine version. Conditions were applied in an ABA design. We counterbalanced across the sample, which pairing of condition and version was presented first.

Capuchin Monkeys. The capuchin monkeys received four sessions. The first three sessions followed the design and counterbalancing applied to the chimpanzees. Due to problems with the foraging version in this group, we ran a second foraging version in a fourth session where partitions in the foraging buckets prevented rush emptying of the entire content. Continuing the alternating condition order (ABAB), half of the capuchin monkeys received in the fourth session the same condition as in the first foraging version, while the other half received a condition opposite to the one in their first foraging session.

2.1.5. Computational model

To predict the behavior of an ideal learner capable of multiple levels of abstraction, we adapted the Probabilistic Hierarchical Bayesian model introduced by Kemp et al. (2007). While Hierarchical Bayesian models have successfully characterized a variety of human adult and child behavior, Felsche and colleagues (2023) were the first to directly test this model of overhypothesis formation in children and non-human primates. Here we apply their model to the current task design, including the variation of only one item dimension (type) and the idea of switching from a current container to another one instead of making a binary choice between containers. Similar to previous work on optimal foraging and utility-based optimal decision-making (Cain, Vul, Clark, & Mitroff, 2012; Jara-Ettinger, Schulz, & Tenenbaum, 2020; Lucas, Bridgers, et al., 2014; McNamara, 1982; Olsson, Brown, & J., 2006), we assume that, while the learner updates their representations of the containers with each new piece of data, the learner's primary goal in deciding whether or not to switch containers is to probabilistically maximize their expected utility (versus other possible goals such as maximizing information gain). As discussed in more detail below, participants stay at the first container as long as its expected value is greater than that of the second container. They then start to probabilistically switch containers in proportion to the difference in the container's expected value if the alternative container promises better rewards than the one they are currently sampling from.

We adapted the a priori model predictions to species-specific factors like the amount of received evidence and each species' reward utilities that were inferred based on the results of preference testing. To evaluate each species' ability for abstract knowledge formation, in addition to comparing performance to the predictions of the full HBM, we also compared their performance to simpler alternative models differing in their capability for abstraction at various levels. The HBM model was implemented in WebPPL (Goodman & Stuhlmüller, 2014), while the preference inference model was implemented in R (R Core Team, 2019).

Fig. 2 provides an overview of both our task and the generative computational model of this task–how the model formalizes the way in which evidence is sampled from the containers. The model assumes that reward items $\boldsymbol{y}^{\boldsymbol{i}}$ are randomly sampled from evidence containers $\boldsymbol{i}$, each of which has a specific item distribution (θ^i) of item types (k = 2 types, high- and low-value). These item distributions within containers represent a first level of abstraction (Level 1), capturing e.g., "this container has mostly high-value items", and are described by a multinomial function. The number of samples was adapted to each species' amount of seen evidence (10 items per container for non-humans, 4 or 6 items per container for children (in total 10 high- and 10 low-valued items)).

The model assumes that the containers themselves were sampled from a higher-level distribution. In other words, the per-container item distributions are, in turn, constrained by an overhypothesis at the second level of abstraction (Level 2), capturing e.g., "containers are mostly uniform in type of item", and "overall, there are roughly equal amounts of high- and low-value items". This second level is described by a Dirichlet distribution parameterized by two hyperparameters: α and **β**. α describes the extent to which item types within each container tend to be uniform (e.g., all are the same or an equal mixture of types). **β** describes the overall composition of item types across all containers (e.g. many high-valued and only a few low-value items or an equal amount of both types). The α parameter is sampled from an exponential prior that assumes a fairly uniform distribution across the probabilities for different item compositions in containers. This corresponds to not having a strong a priori belief that would favor one of the contrasted conditions (mixed or uniform) or over the other, ahead of seeing the evidence. We sample **β** from a symmetric Dirichlet distribution which corresponds to a model that does a priori to assume an equal distribution of item types.

With these assumptions about how the observed evidence is generated, the model uses standard Bayesian updating to simultaneously infer the parameters (overhypotheses) defining the item distributions at the first level (within individual containers, θ^i) and second level of abstraction (across containers, α and **β**), given the sampled evidence from the individual containers (see Appendix for further technical details). In the uniform condition, where the learner is presented with items that are uniform in type from each container, α will be updated to anticipate that any new containers will have distributions that are also highly peaked around a single item type (uniform or near uniform) and therefore will contain either uniformly (or nearly uniformly) high or low-valued items. In the mixed condition, α would expect more equal item distributions within containers, expecting novel containers to have a similar probability of sampling high- and low-valued items (see Table S20 and S21 in the SM for the expected container distributions in each condition).

To predict the participant's choice behavior in the test situation, the model first needs to estimate the probability distribution over the items in each test container ($\theta^{i+1}, \theta^{i+2}$), and over the type of the potential next sample from each test container (y^{i+1}, y^{i+2}), using the overhypotheses inferred about containers in general. With every new low-value sample from the first test container, the model updates the estimate of the item distribution for this current container, θ^{i+1}, using both the observed sample and what it has learned about containers in general from the preceding evidence trials (through the updated hyperparameters, α and **β**). In turn, the model also updates the hyperparameters from the sampled evidence in the first test container. Further, the model also estimates the item distribution in the second test container, from which no items have been sampled yet, based on the current values of the hyperparameters. Thus, after each low-value sample from the first test container, the model has an updated probability distribution over what it thinks the next item from the container will be and what it thinks the first item from the second container would be. As in two of our three conditions the number of items is hidden, for simplicity the model samples with replacement from the test containers. Thus, removing a low-valued sample from, e.g. a mixed container will not decrease the probability of receiving low-valued samples in the future but rather the opposite.

To predict the actual switching behavior, we added a rational switching rule drawn from the psychological literature (Luce-Shepard choice rule; Luce, 1959; Shepard, 1957) to the original model structure by Kemp et al. (2007). We assume that the expected utility of a container is calculated by summing the utilities of each item type **u**, weighted by that item's probability of being the next sampled type (Lucas et al., 2014). Since sampling is assumed to be random, this is equivalent to weighting by the inferred distribution of item types within that

container. Naturally, disengaging from a current activity (sampling from container 1) and switching to another (sampling from container 2) involves some cost. However, we assume this cost to be small in our study as containers were placed relatively close together and switching was not particularly spatially or temporally effortful; thus, it might primarily involve cognitive effort like attention shifting (see General Discussion). Consequently, we assume that a learner's probability of staying with the current container is proportional to its expected utility and the expected utility of the alternative following the Luce-Shepard choice rule (Luce, 1959; Shepard, 1957; Swait & Marley, 2013 also used in, e.g. Jara-Ettinger et al., 2020; Lucas, Bridgers, et al., 2014). The resulting switching probability is 1- the probability of staying (see Appendix). Using this switching rule, we assume that a participant will not switch when the next possibly sampled item from the first container is of greater or equal value to the one sampled from the second container. This is broadly consistent with models in the human and animal optimal foraging literature, which suggest that foragers should consider leaving when the predicted value of staying is less than the predicted value of searching a new location (e.g., Cain et al., 2012; McNamara, 1982; Olsson et al., 2006).

Switching probabilistically occurs when the alternative container promises greater value, with a switch rate proportional to the difference in the expected value of the next sample from each container. Thus, the probability of switching increases with the growing utility advantage of the second over the first container. However, the model never predicts a 100% switching rate, even if the second container promises a certain high-value item and the first container a certain low-value item. This is because the high-value items are not infinitely more preferred than the low-value items, but only to a certain degree as inferred from the preference testing (see below) and participants are expected to choose proportional to the value difference between the containers. This reluctance to switch in the case of equally valued containers can also be seen as capturing the small cost of switching containers in our experiment. However, to further analyze the potential role of assumed switch costs in the participants' behavior, we conducted an exploratory analysis of an added switch cost parameter. The switch cost was established as a constant subtracted within the switching rule (see SM for details). We conducted a parameter sweep and determined the assumed switch cost for each species that best described the data.

To obtain the utilities for each reward type (i.e. the relative value of high and low-valued rewards to each species), we used the preference inference model developed by Lucas, Bridgers, et al. (2014), see Appendix) for inferring overall relative item utilities from a two-alternative forced-choice preference testing procedure. Preference testing was conducted for children (see SM) and capuchin monkeys (Felsche et al., 2023). Unfortunately, due to time constraints, we could not conduct preference testing with the chimpanzees. Here, we used the monkeys' preference values for high- vs low-value items as an approximation for chimpanzees' relative utilities, as initial high and low-valued categorizations for both species were based on the caregiver's estimations. During the data collection, we later observed that chimpanzees consumed presumably low-valued items to a lesser extent than the high-valued food types. However, the chimpanzees consumed relatively more low-valued items than the capuchin monkeys, who almost always refused them. Thus, the monkey's preference values might slightly overestimate the chimpanzees' actual preferences.

2.1.6. Model predictions

Based on the preference testing with children and monkeys, we inferred large utility differences between high- and low-valued items for both species (children:Δ1.50; monkeys (and chimpanzees): Δ1.56). As described above, we used these values to make a priori predictions about the expected switching behavior. We calculated the probability that a learner would switch away from the first test container for each possible number of sampled items from the first container (1 to 10). This represents the normative probability of switching after each sampled item, given the modeling assumptions described above, and that species' inferred item preferences, thus providing a baseline against which to evaluate participant performance. Then, we determined the difference between conditions by subtracting the predicted switching probability at each sampling event in the mixed condition from that in the uniform condition.

As expected, the full idealized model described above, based on Level 2 abstraction (abstraction across containers), predicted that in the test situation, participants should switch earlier (after seeing fewer items) in the uniform as compared to the mixed condition (see Figs. 3 and 4). For example, after sampling the first test item in the uniform condition, the model estimated a 91% probability for a subsequent low-value item from this source and a corresponding 57% probability for a next low-valued item from the second container[2] (see Table S20 in the SM). Based on the reward utilities and the proportional switching rule described above, the Level 2 model thus predicts that around 40% of the participants should switch immediately after the first sampled item in the uniform condition, with a slightly increasing switching rate for the remaining participants after each of the following samples (see Fig. S6 in the SM). In the mixed condition, only around 15% of participants are predicted to switch after the first test sample (the probability for the next item to be of low value is 64% for the first container and 53% for the second container). As the prognosis for the first container gets progressively worse, the more low-valued items are sampled from it, and the estimate for the alternative container remains relatively constant (see Table S20), the switching rate is predicted to increase after each sample. For both conditions, the model predicts that after sampling the 9th item, almost all participants should have switched to the second container (children: mixed = 95.6%, uniform = 99.5%; non-human primates: mixed = 96.1%; uniform = 99.6%; see Fig. 4). Importantly, this confirms that, in principle, an overhypothesis that licenses large differences in switching rates is learnable from the amount of evidence and a priori item-type utilities presented in this study. We also formalized a lesioned model capable of only Level 1 abstraction (abstraction from sampled items to the specific container they were sampled from). This model was solely informed by the low-valued samples from the first test container and the fixed priors for the hyperparameters while ignoring the preceding evidence from other containers. This results in the same prediction for both conditions. With accumulating low-value samples from the first test container, the model assumes an increasing majority of low-value items in this container (e.g. after 1 sample, 60% low-value items, after 5 samples, 78% and after 9 samples, 85%). However, as the model does not update the hyperparameters, the predictions for the second test container stay at the level of the priors, assuming an equal chance of a low and high-value item, independent of the number of samples from the first container. As a result, after every new sample from container 1, between 14 and 43% of the participants that have not yet switched are expected to do so. To account for other potential switching strategies that do not involve abstract reasoning, we compared participants' performance to two random or heuristic learners that operate on different switching strategies. As in the Level 1 model, the random models do not consider condition-specific evidence; thus, they predict no difference in switching rates between conditions. In the first random model, the learner's decision to switch at each point is at chance level, disregarding all samples. Thus, 50% of the remaining participants switch after each new sample from the first test container (Chance 1). This leads to an immediate switching rate after the first item of 50% and a prediction that over 90% of participants should have

[2] The predictions for the unsampled second container are not at 50%, because the low-value samples from the first test container also update the overhypothesis hyperparameter **β**, suggesting that there may be slightly more low- compared to high-valued items in the environment. In the uniform condition this would translate to predicting more uniformly low-value than uniformly high-value containers.

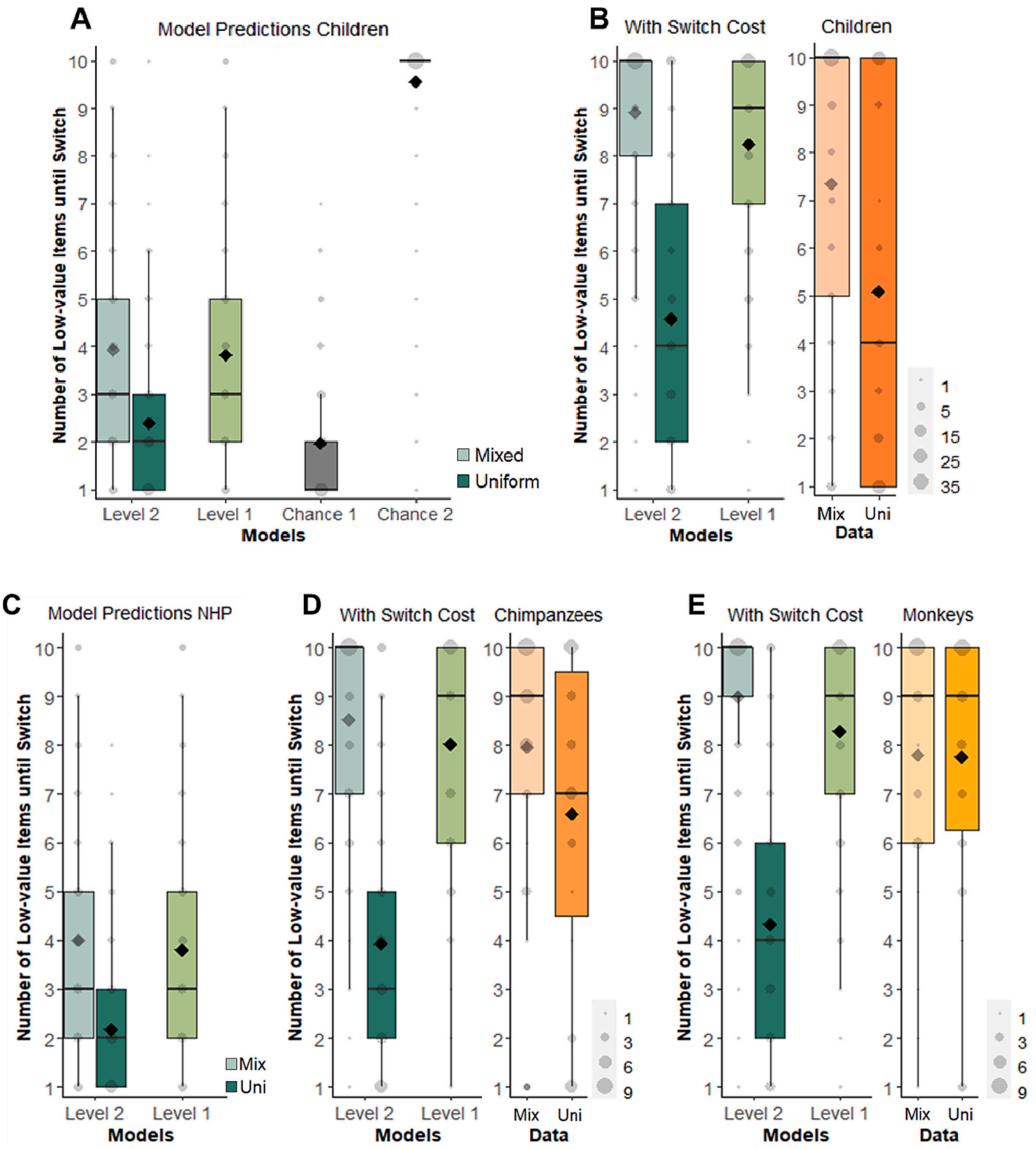

Fig. 3. Model predictions and empirical data for the mean number of low-valued items sampled from the first container before participants switch. Means are indicated by diamond symbols, medians are indicated by horizontal lines. A: Predictions from the full model capable of Level 2 abstraction (across containers), a lesioned model only capable of Level 1 abstraction (from samples to container population and vice versa), a Chance (1) model predicting that after every sample half of the participants switch and a Chance 2 model assuming random contents in containers, all adapted for children's preferences. The predictions in the form of probabilities were multiplied by the factor 100 to simulate a study with 100 participants in each condition. B: Model predictions for the Level 2 and Level 1 model with the switching costs (left) that best matched the empirical child data (right) for children. C: The same predictions as in A adapted to the monkeys' preferences (the chance predictions are identical to A and thus not depicted). D) Same as in B for the chimpanzee data E) Same as in B for the capuchin monkey data.

switched after four samples (see Fig. 4). The second random model (Chance 2), formalizes a learner who randomly predicts the value of the next sample from each container or assumes that, in general, containers have an equal 50/50 mix of high and low-valued items. Based on our switching rule, equal container utilities cause the learner to always stay with the current container. Thus, we introduced a small error term so that 1% of the participants would switch despite assuming equal item distributions in containers (Chance 2). This leads to a very low switching

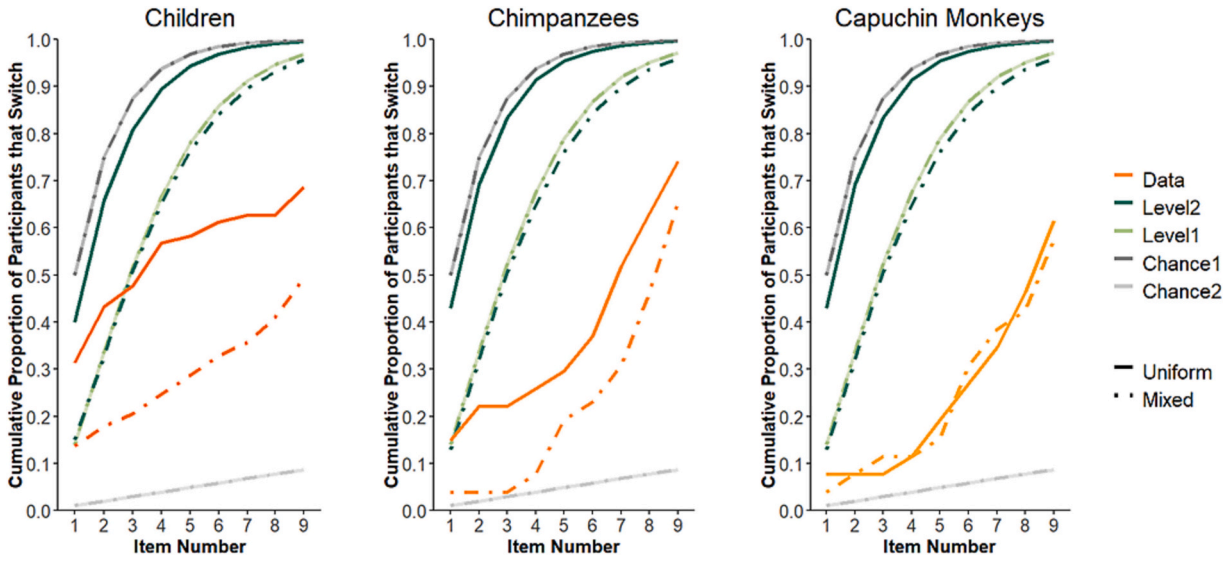

Fig. 4. Model predictions and empirical data for each species, showing the cumulative switch rates after each low-valued sample from the first test container. A difference in switching rate between conditions is seen in children and, to a lesser extent, chimpanzees, but not capuchins. Only the level 2 model predicts a difference in switching rate between conditions. In contrast, the level 1 model and both chance models did not consider condition-specific evidence and thus made identical predictions for both conditions. The chance 1 prediction reflect that 50% of the remaining participants switch after every sample. The chance 2 prediction assumes a minimal error switching rate of 1% when participants predict item types in each container randomly.

rate so that after the 9th item, <9% of the sample has switched.

2.2. Scoring and analysis

2.2.1. Dependent variables

Number of Items Until Switch. To determine the amount of evidence a learner has received during the test phase before they switch, we measured the number of items sampled from the first test container before switching. This variable precisely captures the received information and thus is compared to the model predictions. However, as this measure is not suitable for the foraging version in chimpanzees and children (as here it is unclear how many items participants felt inside the filling material), we did not include it in the statistical analyses for all versions (but see analysis in the SM with this variable for the cup and machine versions and the second foraging version for capuchin monkeys).

Time Until Switch. As an approximation to the number of items participants have received in each version, we determined the time in seconds until the participants switched from the first test container to the alternative. The time started as soon as the participants touched the filling material/the machine/a cup (depending on the version) of container A and ended as soon as they touched the filling material/the machine/a cup of container B. If a participant indicated the wish to leave or stopped interacting with the test materials for one minute, the time ended after the last interaction with the materials. We analyzed this variable because the time until the switch could be measured in all versions and species. Individual variations in sampling speed were expected to be random and to not vary with the experienced condition.

2.2.2. Statistical analysis

All sessions were video recorded. A second observer naïve to the experiment's hypothesis coded 20% of all sessions. As the number and the time variable are measured on an interval scale, we used the Pearson correlation for comparisons between observers. Interrater-reliability was very good for the number variable in the cup and machine version (children: $r(23) = 0.99$, $p < 0.001$; chimpanzees: $r(7) = 0.83$, $p = 0.005$; capuchin monkeys: $r(7) = 0.99$, $p < 0.001$); and for the time variable in all versions (children: $r(41) = 0.94$, $p < 0.001$; chimpanzees: $r(15) = 0.99$, $p < 0.001$; capuchin monkeys: $r(14) = 0.85$, $p < 0.001$).

All statistical analyses were conducted in R (R Core Team, 2012) using the packages lme4 (Bates, Mächler, Bolker, & Walker, 2015) and emmeans (Lenth, Singmann, Love, Buerkner, & Herve, 2018). All presented analyses with chimpanzees and capuchin monkeys were preregistered (https://osf.io/r29nw/?view_only=72d0a6be37fd4f4ca674da847def3181, https://osf.io/prhgq/?view_only=60c206040a5f4132ada70a5d07f6bf35). To answer for each species whether the participants formed abstractions based on the sampled evidence, we examined whether there was a difference between the mixed and the uniform condition in the time until participants switched away from the first test container. As pre-registered, we were interested in whether the condition effect varied depending on the presented version (foraging, cups, machine). Instead of running multiple independent *t*-tests for each presentation version we ran equivalent paired contrasts on our regression model, requiring fewer independent statistical tests (for results of the registered independent t-tests analysis are consistent, and are presented in the SM). In addition, some secondary pre-registered analyses were not run due to insufficient data (see SM for further explanation).

Children. We used a linear model to analyze possible interactions between the versions and conditions and main effects for these factors across the whole sample. For the children, we were further interested in a possible developmental effect and thus included a three-way interaction of version, condition, and age (continuous and centered) in the model. We used the box-cox method by applying the function powerTransform from the R package car to all linear model analyses to account for failed assumptions of normality. The model outputs were analyzed using a type 3 ANOVA based on *F* tests. In case of a significant three-way interaction, we used the emtrends function of the emmeans package to see how the condition contrast changes with age depending on the version. To analyze the difference between conditions separately for each version, we conducted pairwise comparisons using the emmeans function.

Non-human Primates. Due to the within-subject design for chimpanzees and capuchin monkeys, where some subjects did not complete all versions, we used linear mixed-effects models (function lmer) for these species. We conducted two linear mixed effects models, one for each non-human species. In these models, we included a condition-by-version interaction. Due to the high overlap between the factors of session and condition (for each session-level, only one to two condition

levels), we did not include the session factor in the analysis. However, the random effect of participant was included in the model to account for individual effects in the repeated testing. As in children, the model outputs were analyzed using type 3 ANOVAS based on *F* tests. Again, we conducted pairwise comparisons of the conditions within each version using the emmeans function. For the capuchin monkeys, the data from the first foraging version was unreliable since they removed most of the evidence items in their first action on the container. As a result, their switching behavior could not be adjusted to an accumulating amount of witnessed evidence. Thus, we excluded the data from the first foraging version from the analysis and instead only used the data from the second foraging version (Foraging 2) for all analyses with the capuchin monkeys.

2.2.3. Model comparison

Because the number of items sampled before the switch could not be reliably determined in the foraging version of chimpanzees and children (as participants may have sampled items tactilely and did often not remove the low-values items from the containers), only data from the cup and machine versions and the second foraging version in capuchin monkeys was used for all model comparisons. Our main aim was to analyze whether participants of each species showed a difference between conditions consistent with the predictions of an idealized learner capable of Level 2 abstraction. Thus, we correlated the model's predicted condition difference in switch rates after each sampled item with the corresponding empirical difference in switch rates. To determine the empirical switching rates, we calculated for each condition and test item number, the percentage of remaining participants still sampling from the first container, who switched after each item (e.g. 15 participants sample the 6th item from the first test container, 3 of them switch, thus, the switching rate after the 6th item is 3/15 or 20%). To compare the model predictions, which assumed an unlimited number of possible samples, to the observed data, where sampling ended with the tenth item, only the switch rates for items 1–9 were considered because all remaining participants by task design had to switch after sampling the 10th item. As the lesioned model only capable of Level 1 abstraction and the two chance models do not take any condition-specific evidence into account, their condition difference is zero. Thus, no correlation can be computed for these alternative models.

Further, we compared the absolute switching rates of the model and the empirical data. Here, we were able to compare both the Level 2 model predictions as well as those of the lesioned model (Level 1) and both chance predictions (Chance 1: random decision to switch, Chance 2: random prediction of next sampled item) to the children's, capuchin monkeys' and chimpanzees' behavior. To evaluate how well the model predictions describe the data, we compared them separately for each species by minimizing the negative log-likelihood for the model's prediction of the observed data. In this process, empirical switch rates and model predictions for each of the four models were compared separately for every trial and condition before calculating a sum score for each model and species. Model comparisons were then conducted using the difference in AIC scores. AIC scores determine model fit while favoring simpler models with fewer free parameters. $\Delta AIC > 2$ is generally considered strong support for the higher-scoring model. In an exploratory analysis we conducted a parameter sweep for potential switching costs from 0 to 0.5 in steps of 0.01 for the Level 2 and Level 1 model (across both conditions) and compared them to the data using the differences in AIC scores. In a final step, we compared the Level 2 and Level 1 model with their respective switch cost value that best described the data.[3]

[3] The exploratory analysis of switch costs was recommended by a reviewer.

2.3. Results

2.3.1. Statistical results

Children. The linear model revealed a highly significant effect of condition ($F(1) = 20.13$, $p < 0.001$), showing that overall, children switched sooner in the uniform (M = 32.73 s) than in the mixed condition (M = 46.92 s; see Fig. 3). The main effects of version ($p = 0.055$) and age ($p = 0.065$) were trending towards significance (see Table S4). Children tended to take the longest time to switch in the cup version (M = 45.16 s), followed by the foraging (40.89 s) and machine version (33.80s).

The condition by age interaction trended towards significance ($F(1) = 3.74$, $p = 0.054$), showing that overall younger children tended to differentiate more between conditions than older children (Fig. 5, right). Looking at the plot for age in years (Fig. 5, left), 3-year-olds show a strong difference between conditions in the cup and foraging version but do not seem to consider the evidence in the machine version. Four-year-olds show a more consistent difference between conditions across versions. In contrast, 5-year-olds differentiated most clearly between conditions in the machine version but showed only small differences in the other two versions. All other two-way interactions were not significant ($p \geq 0.16$).

There was also a trend towards significance in the three-way interaction of version by condition by age ($F(2) = 2.88$, $p = 0.058$). The emtrends analysis showed that the condition difference was significantly reduced with increasing age in the cup version (estimate = −0.005, SE = 0.002, $t(200) = -2.46$, $p = 0.01$) but not in the other two versions (foraging: estimate = −0.003, SE = 0.002, $t(200) = -1.55$, $p = 0.12$; machine: estimate = 0.001, SE = 0.002, $t(200) = 0.77$, $p = 0.44$). The two-way interaction between version and condition was not significant, however pairwise comparisons of the conditions within each version revealed a strong significant difference in the cup version and foraging version but not in the machine version (see Table S5 and Fig. 5).

Chimpanzees. The linear mixed model (Table S6) revealed a significant condition by version interaction ($F(2) = 3.42$, $p = 0.04$) as well as a significant main effect of condition ($M_{uniform} = 30.14$, $M_{mixed} = 37.74$, $F(1) = 4.61$, $p = 0.04$). The main effect of version trended towards significance ($F(2) = 48.72$, $p = 0.06$). Pairwise comparisons showed that the chimpanzees in the machine version switched significantly quicker in the uniform condition than in the mixed condition (estimate = 26.12, SE = 8.66, $t(60.1) = 3.02$, $p = 0.004$). We found no significant differences between conditions in the other two versions (see Table S7 and Fig. 6).

Capuchin Monkeys. The linear model revealed no significant condition by version interaction ($F(2) = 0.44$, $p = 0.65$). The main effect of condition on the time to switch shows a slight trend towards significance ($F(1) = 2.88$, $p = 0.098$, $M_{uniform} = 22.85$ s, $M_{mixed} = 33.69$ s). In contrast to the chimpanzees and children, where the "number of samples before the switch" variable confirms their sensitivity to the condition-specific evidence (see Table S10 and S12), the capuchin monkeys show no difference between conditions in the number of items they sample before the switch ($M_{uniform} = 7.77$ samples, $M_{mixed} = 7.81$ samples, $p = 0.97$, Table S14). For the capuchin monkeys, the time and number analyses are based on the same data, as in both, all 3 versions were included (as here the foraging 2 version provided clearly countable results). Thus, the marginal condition difference in the time variable is only based on capuchins sampling the test evidence slightly slower in the mixed as compared to the uniform condition but not on an actual difference in the number of samples seen before the switch. Pairwise comparisons (of the time and number variable) showed no significant differences between the uniform and the mixed condition within any version (see Table S9 and S15; Fig. 6). In addition, the model revealed that the effect of version was trending towards significance ($F(2) = 2.99$, $p = 0.06$), as monkeys switched slightly earlier in the cup (M = 18.15 s) compared to the foraging 2 (M = 33.25 s) and machine (M = 35.94 s) version.

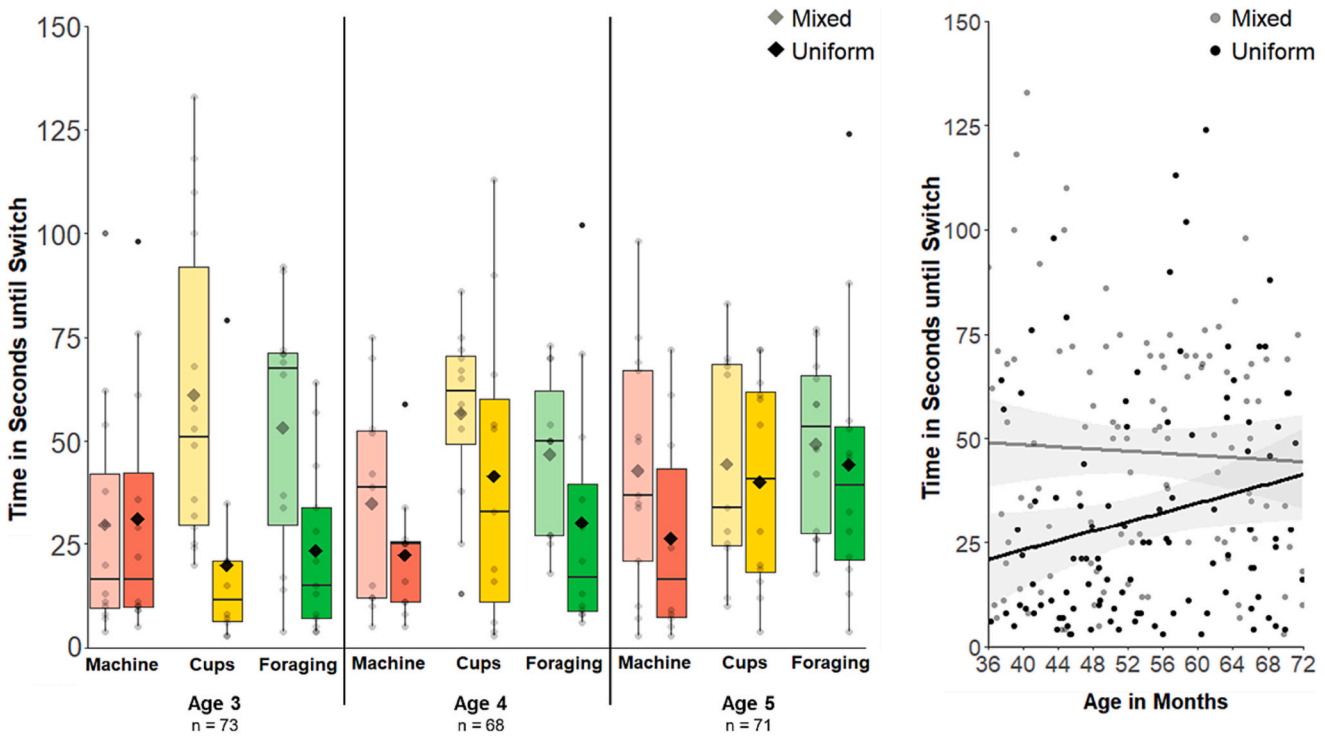

Fig. 5. Results for children by condition and version showing the time in seconds until the switch. Left: Results separated by age group. Means are indicated by diamond symbols, medians are indicated by horizontal lines. Note that we included age as a continuous measure in the analysis and only used age in years here for presentation purposes. Right: Results for the time until the switch dependent on age for both conditions collapsed across all versions. Displayed are the individual data and the regression lines for each condition.

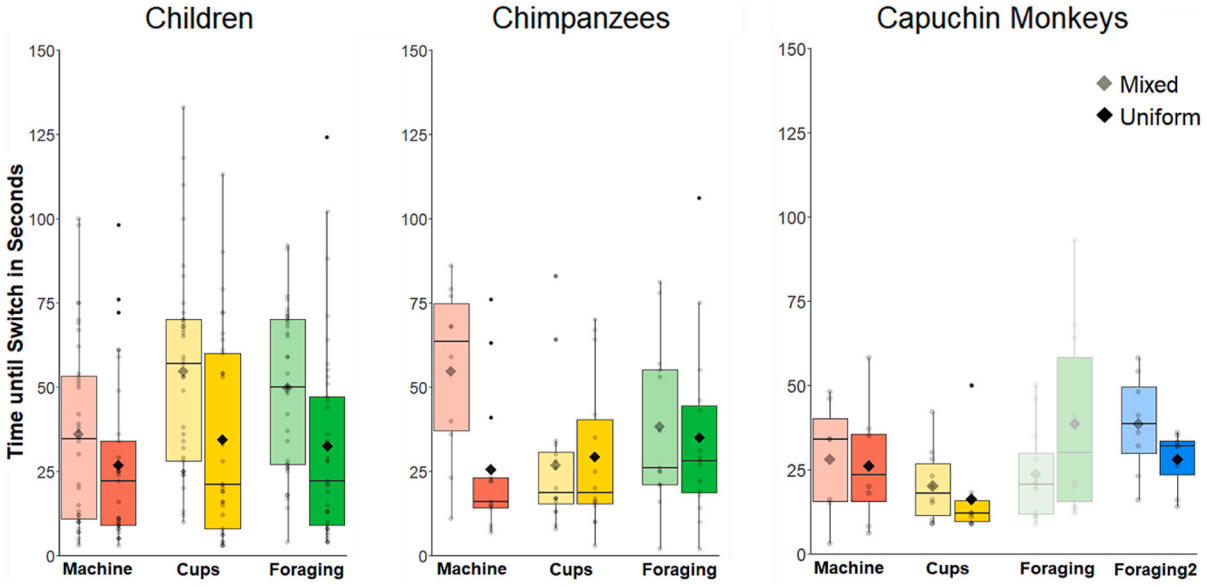

Fig. 6. Results for the time until the switch measured in all species, versions, and conditions. Means are indicated by diamond symbols, medians are indicated by horizontal lines. For the capuchin monkeys, only data from Foraging 2 was used for all analyses due to the unreliability of their first foraging version).

2.3.2. *Model comparison*

The change in the condition difference in children's switching rate across test samples correlates highly with the level 2 model prediction ($r(7) = 0.84$, $p \leq 0.01$; see Fig. S6). Both, children and the Level 2 model predictions show the largest difference in switching rates between the uniform and mixed conditions after the first item, which decreases progressively with subsequent samples. Here, the Level 2 model predicted that, for all species, about 40–43% of participants should switch after the first low-value sample in the uniform condition but that only around 13–15% should switch after one test sample in the mixed condition. The corresponding empirical values are 31% and 14% for children and 15% and 4% for chimpanzees.

The correlation between chimpanzees' predicted and empirical condition difference after each sample is of medium strength but not significant ($r(7) = 0.40$, $p = 0.28$). This is likely because, while the model and the chimpanzees exhibit a parallel decrease in the switching rate difference from the first to the 5th test sample, thereafter the empirical switch rate difference does not decrease as predicted (see

Fig. S6). The capuchin monkeys' empirical condition difference is small and shows no clear direction. Thus, it is not correlated with the level 2 model predictions ($r(7) = -0.07$, $p = 0.86$). Notably, in the non-human species, the number of switching individuals per condition after each new sample is very small ($n \leq 5$). Consequently, the behavior of a single individual can greatly impact the switching rate difference. Thus, the correlation results for these species should be interpreted with caution.

Regarding the absolute rate of switching across the sampled items during the test, the model without any switching costs predicted much quicker switching in both conditions than we saw in all three species. Thus, the empirical data were in absolute terms best predicted by the Chance 2 (children and capuchin monkeys) or Level 1 model (chimpanzee) which predict no condition difference but overall lower switching rates (see Fig. S6, S8 and the SM for more details). While the Level 2 predictions indicate that <3% in both conditions would fully empty the first container before switching, more than a third of each species sampled all ten items from the first container before exploring the alternative (see Fig. 4). This kind of comparative "stickiness" or reluctance to switch relative to an idealized model has been found in previous foraging and patch-switching tasks (Cain et al., 2012; Hutchinson, Wilke, & Todd, 2008). Thus, we considered the possibility that the participants incorporated switching costs into their foraging decision during the test situation. Indeed, the parameter sweep showed that all three species match Level 2 and Level 1 model predictions best when including a switch cost of around 0.35 (see SM for detailed results). When comparing the models with their respective optimal switch cost value, the Level 2 model predicts the children's data best; while the Level 1 model still outperforms the Level 2 model for the non-human species (see Table S26). Likewise, when focussing on the relative switching rates after the first sample only, even without considering switching costs, the Level 2 model matches children's performance best, while chimpanzees and capuchin monkeys are best described by the Level 1 predictions (see SM for AIC values).

2.4. Discussion

The results of Experiment 1 show strong evidence for abstract reasoning abilities in 3- to 5-year-old children who switched away from the first test container earlier in the uniform compared to in the mixed condition. While this effect seemed stronger in younger age groups and showed some variability across versions (see General Discussion), those factors did not reach significance. The children's condition difference in switching behavior was well predicted by a hierarchical Bayesian model capable of Level 2 overhypothesis formation. Similarly, albeit to a lesser extent, the chimpanzees' switching behavior also showed a significant difference between conditions that correlated to a medium extent with the Level 2 model predictions, even if this was mainly driven by only one version of the experiment. In general, all species were much more hesitant to switch than the overhypothesis model predicted (see General discussion), and thus, the original model failed to predict their behavior in absolute terms. However, when including a switch cost parameter in the model, the children's behavior is best described by the Level 2 model while the non-human primates absolute switching numbers better match the Level 1 prediction. In contrast to children and chimpanzees, the capuchin monkeys have failed to show any sensitivity to the condition-specific evidence. This suggests that their abstract reasoning abilities are somewhat reduced, less robust or slower than that of apes and humans. However, slight variations in the experimental procedure or species differences in other cognitive abilities could also contribute to this pattern in results. In experiment 2, we further explore the reasons for the capuchin monkeys' failure in experiment 1 by varying the reward structure to be more similar to that of children.

3. Experiment 2

In the first experiment, one crucial difference between the study design for non-human primates and children lies in the reward items' nature. Whereas children received balls that could be used for a game (high value) and entirely non-functional blocks (low value), the non-human primates received food items of different values. Thus, for children, the two categories of high- and low-valued items were clearly distinguishable based on form and function. While the food rewards for the monkeys and chimpanzees had relatively higher and lower values, there is not necessarily a clear categorical distinction. Further, even food items of low value could be eaten and thus were never non-functional or of zero value. To align the reward structure for children and non-human primates, we conducted a follow-up study with the capuchin monkeys (we could not conduct this study with the chimpanzees due to testing constraints). Here, we presented them with only two types of rewards that differed in function and appearance, as was the case for children. Using this simpler design, we predicted that the monkeys might find it easier to learn the overhypotheses about the reward distribution within and across containers and thus show a difference in their switching behavior between the two conditions.

3.1. Method

3.1.1. Participants

We tested 16 capuchin monkeys at the Living Links research centre, all of which had participated in at least one session of the previous study (6 female, age: M = 8.81 ± 3.51 SD, range 5–18 years). Five additional monkeys were excluded from the analysis due to experimenter error (3) or because they asked to leave the testing area before the final test phase, and their session could not be repeated in the given time for the study (2).

3.1.2. Materials and procedure

The procedure was identical to the previous study except that only two different object types were involved: pieces of grape as high-value items and blue stones as low-value items. The non-edible stones were familiar to the monkeys, as they have been used in previous studies, but they were never associated with positive or negative reinforcement. Due to this modification, the monkeys already saw the same low-value item type used in the test phase during the evidence trials. We only implemented this experiment in the machine version, as it provides the most precise measure for the number of items before the switch. It was also the version in which chimpanzees were most successful in Study 1. The appearance of the machines was altered so that carry-over effects from Experiment 1 were minimized (see SM).

The monkeys were assigned to either the uniform or the mixed condition in a between-subjects design. As this version resembled the previous machine version of the main experiment, we counterbalanced whether monkeys received the same condition as in their previous machine session or if they now experienced a different condition in this version. Of the 8 monkeys in the uniform condition, 4 had received the uniform condition in the previous machine version, whereas 4 had previously experienced that machines contained a mix of items. Of the 8 monkeys in the mixed condition, 3 had received the mixed condition previously in their machine version, whereas 5 had previously experienced the uniform condition with the machines. The first machine condition in Experiment 1 was session 3 for all subjects, and the current study was conducted as session 5 (after the second foraging version).

3.2. Results and discussion

As shown in Fig. 7, the capuchin monkeys switched earlier (after less samples from the first container) compared to experiment 1, and they descriptively showed more sensitivity to the condition-specific evidence. However, a two-sided *t*-test revealed no significant difference ($t(13.99) = 1.32$, $p = 0.21$) between the mixed and uniform conditions regarding the number of sampled items before the switch. The correlation of the predicted (Level 2) and empirical condition difference in the switch rate

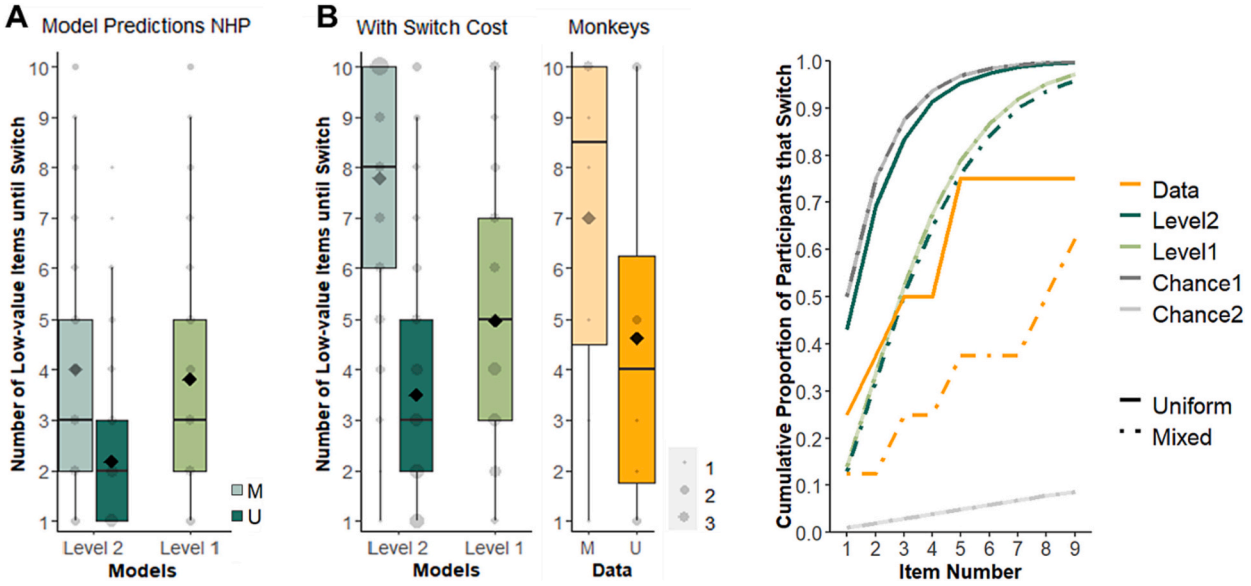

Fig. 7. Results and model predictions for Experiment 2. Left: Mean number of samples before the switch for each condition, including individual monkeys' absolute values. Means are indicated by diamond symbols, medians are indicated by horizontal lines. Right: Empirical cumulative switching rates for capuchin monkeys and the predicted cumulative switching rates for all four modelled learners.

after every test sample is of medium strength but not significant ($r(7) = 0.50$; $p = 0.17$; see Fig. S6 and Model Comparison for details of this analysis). Given the small sample size and the overall low switching rates (per condition and sample number $n \leq 2$ switch), this result must be interpreted with caution. Nevertheless, after the first five samples, 6/8 monkeys in the uniform condition have switched, while only 3/8 did so in the mixed condition, a result that is consistent with the direction of the predicted effect.

Comparing the absolute values for the data and the model predictions regarding the percentage of remaining participants that switch after each sample (without considering switching costs), the Level 1 model (AIC = −83.17) most accurately predicts the capuchin monkeys' choices (comparisons to Level 2 (AIC = −88.35) ΔAIC = 5.18; Chance 1 (AIC = −119.99) ΔAIC = 36.83; Chance 2 (AIC = −110.86) ΔAIC = 17.69). This pattern resembles the results of the children and chimpanzees in Experiment 1, whose results were best described by condition-independent models (Level 1 and Chance 2) with lower predicted switching rates, despite showing a clear empirical condition difference. However, when including the best fitting switch cost parameter to the model, the capuchin's behavior was best predicted by the Level 2 model (AIC = −69.11, see Fig. 3 and SM). Thus, it is possible that the capuchins were able to generalize the condition-specific knowledge to the test situation but also considered the costs of switching containers.

Overall, these findings suggest that reducing cognitive demands by emphasizing category membership and implementing a larger value contrast between categories is a promising route of future study design, but was in this limited sample of capuchin monkeys not sufficient to detect the ability for overhypothesis formation.

4. General discussion

Over the course of two experiments, we examined whether children, chimpanzees, and capuchin monkeys can form overhypotheses about item distributions, given limited evidence, to optimize their search for high-valued rewards in different situations. A probabilistic hierarchical Bayesian model showed that a learner capable of abstract concept formation should switch away early from a container providing low-valued rewards when previously self-sampled evidence suggested that all rewards in a container tend to be the same. However, when previously sampled containers provided a mix of high and low-value items, the model predicted that a learner capable of forming overhypotheses would persist at sampling from such a container for longer before switching to an alternative.

Preschoolers' switching behavior matched the predicted difference between conditions, showing that they can form overhypotheses and generalize abstract patterns across situations already at age 3. There was also tentative evidence for chimpanzees' ability to form abstractions, as they showed the expected difference between conditions in one of three experimental contexts. However, in contrast to the capuchin monkeys, chimpanzees ate many low-valued items during the test. This suggests that chimpanzees may have less extreme food preferences than assumed by the model (which was based on the monkeys' preferences). Thus, the results may represent a conservative estimate of their actual abilities for abstraction. Finally, in contrast to the other species, the capuchin monkeys' behavior in Experiment 1 showed no sensitivity to the condition-specific evidence and was best explained by a model based on random expectations about the item distributions, and thus provided no evidence for abstract concept formation in this species. With more distinct reward categories (Experiment 2) the capuchin monkeys' performance resembled both the model capable of abstractions and children's performance in Experiment 1 more closely. However, the difference in the monkeys' performance across conditions was not statistically significant.

Our results contrast with the usually poor performance of young preschoolers and non-language-trained chimpanzees in abstract relational reasoning tasks (e.g. Christie & Gentner, 2010; Hochmann et al., 2017; Premack, 1983; Walker et al., 2016). Thus, our findings contradict the assumption that flexible abstract reasoning is a human-unique ability that relies on relational language (Gentner, 2003; Penn et al., 2008). As evidence from infants and toddlers suggests, detecting overhypotheses might be present already from an early age – before the emergence of relational language- and possibly supports the efficient knowledge expansion in early childhood (Dewar & Xu, 2010; Sim & Xu, 2015; Walker & Gopnik, 2014). Our results are further supported by another study showing that task designs beyond the RMTS task can reveal abstract reasoning abilities in human preschoolers. In Felsche and colleagues (2023) 4 and 5-year-old children in a passive sampling paradigm were sensitive to the general sorting patterns of reward items based on their type or size and adapted their choices accordingly.

The common critique of studies with non-human participants

claiming that success in RMTS tasks is based on lower-level perceptual processes like comparing the perceptual variability (or entropy) within stimulus pairs (Penn et al., 2008; Wasserman et al., 2017) cannot be applied in our study. Across conditions, we also vary the perceptual similarity of the successively sampled items from each container (e.g. all same vs. different) during the evidence phase. However, we do not believe that a representation in terms of a minimalist version of expected entropy (that does not incorporate some aspect of abstraction) could, on its own, account for success in this task. In the final test phase of our study, participants made choices based on yet unseen, predicted future samples from the test containers instead of being confronted with a forced choice between visible item pairs or arrays. In addition, the test situation in all conditions only presented uniformly low-valued items and children and chimpanzees already showed a condition difference in their switching behavior after only a single sample from the test box. Together with the fact that they never had the opportunity to switch containers in the evidence phase, this eliminates the possibility that the participants learned a simple perceptual rule like: "low perceptual variability of items, switch away from a container; high perceptual variability of items continue sampling, " which would also not be adaptive in the case of high-valued uniform items. In contrast to previous studies showing successful abstract reasoning performance in primates after hundreds or thousands of trials, our paradigm only included a minimal amount of evidence, which further excludes the possibility of a slow, associative learning process contributing to the results.

A concern might be that chimpanzees succeed due to a training effect, as they only showed a significant difference between conditions in the machine version that was always presented last. We believe that this is unlikely as, despite a possibly heightened understanding to pay attention to the reward distribution in the evidence phase, nothing from the previous versions could have been learned that would support a larger condition difference in the third session, as the types of food, presentation and foraging method, and condition changed between versions. In particular, all chimpanzees gained equal experience with mixed and uniform evidence in the first two sessions and thus could not have learned that, in general, across all versions, food items in containers are all mixed or all uniform. Following, they could not expect the distributions found in the final machine version (and indeed, if they had learned something along these lines that would itself be a high-level abstraction). Even if they had learned the association that whenever sweet potato is presented, both test containers will only provide sweet potato (a pattern that is the same in the mixed and uniform version), this would not increase the difference in switching behavior between conditions but rather cause a general motivation decline to engage in the task, which was not the case.

In contrast to those lower-level explanations, our study provides further evidence for more recent arguments explaining population differences in relational reasoning tasks based on context-sensitive inductive biases rather than capacity differences (Carstensen & Frank, 2021; Kroupin & Carey, 2021; Walker et al., 2016). Our task design presented participants with stimuli and overhypotheses that are intrinsically meaningful, an intuitive self-directed foraging mechanism of evidence acquisition, and an ecologically valid pat-switching test situation. Those attributes contrast with the arbitrary stimuli, abstract patterns, and forced-choice situations used in traditional tasks. Thus, our procedure may have provided an improved context to measure our participants' abilities for abstraction.

The difference between the capuchin monkeys, whose performance did not differ significantly between conditions, and chimpanzees that showed some sensitivity to the abstract patterns, is in line with earlier studies demonstrating differences between apes and monkeys in abstract reasoning abilities (e.g. Flemming & Kennedy, 2011; Kennedy & Fragaszy, 2008; Thompson & Oden, 2000). In addition to this capacity difference between the species, variation in related abilities or other skills involved in the task performance are conceivable. For example, monkeys' inability to analyze the overall structure of the evidence could be rooted in a narrowed focus on individual food items. Research on hierarchical stimulus perception supports this assumption. Here, monkeys seem to process the local components with more ease than the global pattern (e.g. De Lillo, Spinozzi, Truppa, & Naylor, 2005; Spinozzi, De Lillo, & Truppa, 2003), whereas chimpanzees show more mixed results (Fagot & Tomonaga, 1999; Hopkins & Washburn, 2002). Another factor contributing to the species difference could be chimpanzees' superiority compared to capuchin monkeys when judging sequentially presented item quantities (Beran, 2001, 2004; Evans, Beran, Harris, & Rice, 2009). In the current study, monkeys and chimpanzees usually consumed at least high-value items as soon as they found them. Thus, they were required to keep track of and summarize the (previously consumed) evidence to detect the pattern underlying the samples. It can be excluded that capuchin monkeys were not motivated to engage in efficient search in an experimental context (De Lillo, Visalberghi, & Aversano, 1997). Further, it is unlikely that capuchins were insufficiently motivated to look for more high-value food in the test situation after eating the evidence items. The overall food amount was equal to amounts used in previous studies conducted with these same monkeys, in which no motivational decline was observed, and the monkeys' relative preference for the high- vs. low-value items was high (e.g. Tecwyn et al., 2017). As mentioned earlier, the less extreme food preferences of the chimpanzees as compared to the capuchin monkeys potentially led to an underestimation of the actual species difference in abstraction abilities, as it may have reduced the chimpanzees' motivation to look for high-value food after encountering the low-value samples. Supporting the importance of relative food preferences for the chimpanzees' test behavior, their switch behavior is associated with the number of consumed low-valued reward items during the test situation. In all sessions where chimpanzees switched after sampling 5 or less items (and the amount of the low-valued samples during the test could be reliably counted, 13 sessions), they also refused to eat any of the items (only in one session a single food item was eaten (3% of uncovered items)). However, when switching late from the first container, after sampling 6 or more items, the chimpanzees had consumed 39% of the uncovered items (in 21 of 40 sessions). Thus, when chimpanzees consumed more low-valued items, they also sampled more before switching to the next container in search of potentially higher-valued items. In contrast, capuchin monkeys never consumed any of the low-valued test items but still switched relatively late.

In humans, the literature usually shows an improvement in abstract reasoning abilities across the preschool ages from 3 to 5 (e.g. Christie & Gentner, 2010; Christie & Gentner, 2014; Hochmann et al., 2017). However, our study did not confirm this pattern and instead pointed towards a negative developmental trend. This age effect was dependent on the version and primarily based on older children switching later in the uniform condition than younger children, making the condition difference less pronounced in older preschoolers. A study by Ruggeri, Swaboda, Sim, and Gopnik (2019) showed an age effect similar to that in our study. In contrast to 3- and 4-year-olds, 5-year-olds in that study did not connect the knowledge they gained in an evidence phase to a subsequent search situation unless it was made explicit by reminding the children of the events seen in the evidence right before the test phase. One possible explanation for both sets of findings is that older children and adults have stronger prior assumptions about general rules (e.g. that individual items or properties have causal power) and thus might be less flexible in learning or generating new and unusual task-specific patterns based on limited evidence (e.g. that relational features are causally effective; Bramley & Xu, 2023; Gopnik, Griffiths, & Lucas, 2015; Lucas, Bridgers, et al., 2014;). For instance, older children may have a strong prior belief that items tend to be sorted by type rather than mixed together (or vice versa), which the evidence in our study is insufficient to overcome. A more practical explanation could be that older children were less motivated than younger children to repeatedly obtain balls for the marble run, a tendency reflected in the preference testing results (see

Felsche et al., 2023). Thus, they might have been more invested in exploring the containers rather than obtaining high-value rewards as efficiently as possible. Older children might have also formed a stronger normative motivation to adhere to an unintended "game rule" first to empty a container or explore it for a while before being "allowed" to move on to the next container. We tried not to induce such a normative motivation by allowing variation in the number of sampled items per evidence container while avoiding complete emptying. The experimenter also left the immediate test situation to reduce possible social normative pressure. However, these measures cannot entirely exclude the possibility that especially older children have formed a normative motivation in this game-like experimental setting.

Our study showed that the a priori predictions of a hierarchical Bayesian model provide a useful normative model of processes at play early in human ontogeny when extracting abstract patterns and making predictions in new situations). Children, and to a lesser extent chimpanzees, showed a significant difference between conditions in switching rates, a pattern predicted only by the Level 2 model, and not by any of the lesioned alternative models, and the relative difference in switch rate across trials showed a high correlation between the model predictions and children's performance and moderate (but non-significant) correlation to chimpanzee performance.

However, when comparing the absolute model predictions of how many participants should switch after each item to the data, the best overall prediction for children's and chimpanzees' performance was achieved by Chance and Level 1 models that ignore the condition-specific evidence, despite these models predicting no condition difference in switch rates. A likely explanation for this counterintuitive finding is that all species were much more reluctant to switch than predicted by the model. The phenomenon of more conservative switching rates in the empirical data as compared to ideal learner model predictions has previously appeared in computerized search tasks with humans (Cain et al., 2012; Hutchinson et al., 2008) and optimal foraging theory assessments in multiple other animal species (Nonacs, 2001). Hutchinson et al. (2008) argued that optimal models do not consider other factors that influence individuals' behavior like additional intentions (e.g. mate search), risks, and uncertainties (e.g., predator behavior, nutritional state; Nonacs, 2001), as well as a possible alternative motivation to learn more about the environment, which might favor 'stickiness' at the first foraging location.

Consistent with this interpretation, we conducted an exploratory analysis to investigate potential switching costs that may have impacted participants' decision to switch. Here, we found that when including a moderate switching cost in the model, children's absolute rate of switching (and that of capuchin monkeys in experiment 2) was best explained by a Level 2 learner. While the chimpanzees' behavior was still best explained by the Level 1 model, both models matched the apes' and the capuchin monkey's behavior better when including the switch cost parameter. This parameter could represent the loss in time or energy to physically move to the second container. Although the spatial distance between the test containers in our study was minimal, previous studies have also shown that primates engage in spatial and temporal discounting, which might reduce the relative value of the second container (Hopper, Kurtycz, Ross, & Bonnie, 2015; Kralik & Sampson, 2012). Further, anecdotally, we observed that, given that sampling from a container was fast and took little effort, the time needed to switch containers could instead be sufficient to sample at least one more item from the first container. Further switching costs could include the cognitive effort needed to inhibit sampling from the current reward source and shift attentional focus to the next container.

In addition to the switch cost analysis, we also explored other likely factors that the normative model missed but which might still influenced the participants' switching behavior, like the motivation to adhere to an implicit game rule of wanting to empty a container before switching (see SM). It is also possible that the participants acted based on a different prior, for instance, initially assuming that items within containers are more likely to be somewhat mixed. Other possible causes of the comparatively slower switching rate relative to the optimal model across all species include a possible motivation to increase the quantity rather than the quality of acquired items, combined with a learned pattern that once they abandon a container, it becomes unavailable. Thus, searching exhaustively before switching maximizes the overall reward quantity. Further, studies on the endowment effect show that both humans and non-human primates value items more highly once they are in their possession and thus do not always follow the predictions of rational choice models (e.g. Brosnan et al., 2007; Lakshminaryanan, Keith Chen, & Santos, 2008).

Another possibility is that sampling behavior was influenced not only by the expected reward value but also by the informational value of the next sample from each container. Especially given the absence of time pressure and direct observation during the test phase, participants might have sampled longer from the first container to gather more information about its contents and the general item distribution pattern. However, at least for children, it was emphasized that these were the last two containers, eliminating the utility of such information for future searches. While mere curiosity about the first container's contents is a possibility that could have led to later switching, the first item from the second container with completely unknown content holds even higher informational value, which should have motivated overall faster switching rates.

Importantly, once an overhypothesis is established, the information value of the sampled items differs by condition: the first item out of the uniform containers provides all of the information about its contents, diminishing the informative value of subsequent items; in mixed containers items are generally less informative. Although participants may have considered informational value, its impact on sampling behavior can be expected to be low, given the overall low empirical switch rates and the non-human primates' strong motivation to obtain high-value items with comparatively lower curiosity rates (Sánchez-Amaro & Rossano, 2023; Forss & Willems, 2022; but see Wang & Hayden, 2019). Future research could include more of these factors (e.g. varying priors, switch costs, normative expectations, informational value) a priori in the model and then generate predictions to manipulate them experimentally to see which role they play in the participants' decision-making alongside overhypothesis formation.

We used multiple versions of our experimental paradigm to examine how generalizable the participants' performance was across different contexts and to maximize each species' opportunities to show evidence of abstract learning if versions were not equally accessible to each species. Our aim was for the different modes of presentation to vary in their optic and haptic properties but not in the presented amount or type of evidence or other factors that could influence the switching rates (e.g. cost of switching). However, one interesting difference between the foraging and cup versions on one side and the machine version on the other is that in the first two conditions, participants could anticipate the total amount of items inside the containers (either by seeing the total number of cups or by successively removing filling material from the foraging containers). When operating the machines, participants never knew how many items were left in the apparatus due to the opacity of the material.

When learners assume small, finite amounts of items in containers, they may anticipate that removing low-valued items from mixed containers increases the chance of retrieving high-valued items later. Consequently, they might remain longer with the first test container in the mixed condition. For the uniform condition, it matters much less whether some low-valued items are removed or replaced after sampling, as the remaining items are highly likely to be low-value either way. Thus, the relative prediction for switching in the conditions (switching later in the mixed condition) stays the same. However, our participants switched later than the model predicted in both conditions, which would not be predicted by assuming finite rewards. Nevertheless, it would be interesting to explore the effects of the assumed size of the item

populations in containers and, thus, their level of depletion on the search behavior and the speed of forming abstract concepts. Expanding the model with such a variable would help to explain performance differences between versions. The multi-version design reveals the context-dependency of participants' spontaneous ability to form abstract concepts, which is also evident in children's variability in RMTS performance (e.g. Christie & Gentner, 2014; Goddu et al., 2020; Walker et al., 2020). Thus, our study shows that it is crucial for the interpretation of results to implement designs in varying contexts, as only using one version might have led to overly simplistic or even drastically different conclusions. The importance of the testing context is also highlighted by the capuchin monkeys' performance in the second experiment, in which the observed condition difference resembled the predictions of the computational model more closely than their performance in Experiment 1. The change in reward types from high- vs. low-value in experiment 1 to edible vs. non-edible in experiment two has potentially helped this species to differentiate between conditions. However, given the missing statistical significance of the result in this relatively small sample, it is still unclear whether capuchin monkeys can spontaneously form abstract concepts. Nevertheless, these results should inspire future research investigating abstract reasoning abilities in monkeys to move away from highly arbitrary computerized tasks, like the RMTS procedure in which capuchins have only showed success after lengthy training (e.g. Truppa et al., 2011) as well as other procedures with high demands on memory and inhibition (e.g. Felsche et al., 2023) towards more ecologically valid and simplified situations like the patch switching scenario in our study. Further, our study shows that the replication of task paradigms in varying contexts and with multiple materials or reward types seems to be crucial if we want to understand the robustness and nature of different species' abstract reasoning abilities.

In conclusion, our study strengthens the view that the ability for abstract reasoning is present early on in human development and can be characterized by probabilistic hierarchical Bayesian models. Variability in children's performance across tasks and age groups seems to be caused by contextual factors that appeal to learned biases and additional task demands rather than differences in their capacity for abstraction (Hoyos et al., 2016; Kroupin & Carey, 2021). In contrast to previous views arguing for a stark divide between humans and other animals (Gentner, 2003; Penn et al., 2008), or apes and monkeys (Thompson & Oden, 2000) in abstract reasoning, our study supports a perspective of more gradual differences of this ability between species (Carstensen & Frank, 2021; Gentner et al., 2021; Katz, Wright, & Bodily, 2007; Seed, Hanus, & Call, 2011). The study highlights the importance of a multi-version experimental design, especially in a comparative setting, as different species might have different requirements to reveal optimal performance. Further, it shows that applying probabilistic hierarchical Bayesian models produces a more informative species comparison as it allows to incorporate group-specific factors like the received amount of evidence or reward preferences. In addition, the models provide a more comprehensive and testable formalization of the assumed underlying cognitive structure assumed to play a role in different groups of participants. Future studies could extend the approach of using computational models and more variable paradigms to study abstract relational reasoning, perhaps using a wider variety of probabilities between fully mixed and fully uniform. The method used here could be applied across a wide range of species, age groups, and cultures and we suggest it is a promising direction for future work on the origins of abstract thinking.

CRediT authorship contribution statement

Elisa Felsche: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Christoph J. Völter:** Data curation, Formal analysis, Investigation, Writing – review & editing. **Esther Herrmann:** Project administration, Resources, Writing – review & editing. **Amanda M. Seed:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Resources, Software, Supervision, Writing – review & editing. **Daphna Buchsbaum:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing.

Data availability

The data, WebPPL and R scripts associated with this paper are available on OSF (https://osf.io/u9vbp/).

Acknowledgments

We thank Jessica Da Cunha, Nishat Kazi, Katrina Palad, Mrinal Anagal, Justine Biado, Kiah Caneira, and Kay Otsubo for help with the children's data collection. We also thank the Royal Ontario Museum and Ontario Science Centre for hosting this research. We are grateful to the Royal Zoological Society of Scotland (RZSS) and the University of St Andrews for core financial support to the RZSS Edinburgh Zoo's Living Links Research Centre, where this project was carried out. We thank the RZSS keeping and veterinary staff for their care of animals and technical support during data collection. We are thankful to Richard Vigne, Samuel Mutisya, Stephen Ngulu, the board members, and the Sweetwaters Chimpanzee Sanctuary staff in Kenya for their support. 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].

Author Note. We have no conflicts of interest to disclose.

Appendix A. Appendix

Following the model of Kemp et al., 2007 we use a Dirichlet-multinomial model (Gelman, Carlin, Stern, & Rubin, 2014), which describes the relationships between the data and parameters at different levels of abstraction.

Formally, the model is described as:

$$\alpha \sim \text{Exponential}(\lambda).$$

$$\boldsymbol{\beta} \sim \text{Dirichlet}(\mathbf{1}).$$

$$\boldsymbol{\theta}^i \sim \text{Dirichlet}(\alpha, \boldsymbol{\beta}).$$

$$\boldsymbol{y}^i \sim \textbf{Multinomial}\left(\boldsymbol{\theta}^i\right).$$

With y_j^i representing the type (e.g., a high-value item) of the jth item sampled from the ith container, and $\boldsymbol{\theta}^i$ representing the distribution of items types within that container, with $\boldsymbol{\theta}_k^i$ indicating the probability of sampling item type k (e.g., a low-value item) from container *i*. Throughout the main

analyses in this article $\lambda = 0.5$, which corresponds to a fairly uniform distribution across the probabilities for different item compositions in containers. Before seeing any evidence, it slightly favours skewed and uniform distributions over equal mixes of high- and low-valued items but it is sensitive to the evidence presented in both conditions. See the supplementary material for a more detailed analysis of the effect of different values of λ on the model predictions.

The two parameters forming the overhypothesis at the second level of abstraction are α, describing the extent to which item types within each container tend to be uniform and **β,** which describes the overall composition of item types across all containers. To model overhypothesis formation, we infer $p(\alpha, \boldsymbol{\beta} \mid Y)$, the posterior distribution over (α, **β**), given the observed items ***Y***, drawn from the N evidence containers,

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

estimated using the Metropolis-Hastings Algorithm. Here we used 1 chain with 500,000 samples, a lag of 10 and a burn in of 1000.

To predict the participants' behavior in the test situation, the model estimates the expected distribution over item types for the next sampled item j from the first test container by marginalizing $p\left(y_j^{i+1} \mid y_{-j}^{i+1}, \alpha, \beta\right)$ the predicted probability of the next sample from the first test container $i+1$ being of a specific type, given the already known samples from this container, y_{-j}^{i+1} (everything not j), and a specific overhypothesis, represented as the expected value of α and **β**, estimated from the evidence containers,

$$p\left(y_j^{i+1} \mid y_{-j}^{i+1}\right) = \int \int p\left(y_j^{i+1} \mid y_{-j}^{i+1}, \alpha, \beta\right) p(\alpha, \beta \mid Y) d\alpha, d\beta \tag{2}$$

Which we approximate by averaging $p\left(y_j^{i+1} \mid y_{-j}^{i+1}, \alpha, \beta\right)$ across sampled values of $p(\alpha, \beta \mid Y)$. For a Dirichlet-Multinomial distribution, $p\left(y_j^{i+1} \mid y_{-j}^{i+1}, \alpha, \beta\right)$, the posterior predictive distribution for the type of the next item in the container, given a fixed set of hyperparameter values, has a simple known closed form solution.

$$p\left(y_j^{i+1} = k \mid y_{-j}^{i+1}, \alpha, \beta\right) = \frac{N_k + \alpha\beta_k}{\sum_{m=1}^{K} N_m + \alpha\beta_m}$$

The predictions for the type distribution in the second, yet untouched, test container and thus also the predictions for the next sample from this container are inferred in this same manner. However, as there are no observed items from this container, $N = 0$, so the inference is based solely on the updated overhypotheses.

Finally, the choice of whether to continue sampling from the first test container or to switch to the second one is determined based on the expected utility of this container. The expected utility of a container is calculated by summing the utilities of each item type (see below), weighted by its probability of being the next sampled type. Assuming that learners prefer a container proportional to its relative utility (the bigger the difference between containers, the more the expected higher value container is preferred; Luce-Shepard choice rule, Luce, 1959; Shepard, 1957; Swait & Marley, 2013), the probability to switch is calculated:

$$P(c = i \mid u) = 1 - Min\left(1, \frac{e^{u_1}}{e^{u_2}}\right) \tag{3}$$

The relative utilities **u** of the high and low-valued rewards are inferred from the preference testing choices **c** using the preference inference model described in Lucas, Bridgers, et al. (2014). Again, it is assumed that a learner becomes increasingly likely to choose an option as its expected utility increases. However, this choice is treated as a simultaneous choice between multiple items (rather than a choice to switch from the current item to another):

$$P(c = i \mid \boldsymbol{u}) = \frac{e^{u_i}}{\sum_j e^{u_j}} \tag{4}$$

Following Lucas, Bridgers, et al. (2014), we infer item type utilities **u** from learner's choices **c,** by computing the posterior probability $p(c \mid u) \propto p(u \mid c) p(u)$, estimated using the Metropolis-Hastings algorithm. We assume a priori that the type preferences (utilities) are normally distributed, with $\mu = 0$, and variance $\sigma^2 = 2$. Here we used one chain with 10,000 samples and a burn in of 500. We separately inferred type preferences for children and capuchin monkeys, and used the capuchin's preferences as a stand in for those of the chimpanzees, as discussed in the paper.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105721.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Beran, M. J. (2001). Summation and numerousness judgments of sequentially presented sets of items by chimpanzees (pan troglodytes). *Journal of Comparative Psychology, 115*(2), 181.

Beran, M. J. (2004). Chimpanzees (pan troglodytes) respond to nonvisible sets after one-by-one addition and removal of items. *Journal of Comparative Psychology, 118*(1), 25.

Bramley, N. R., & Xu, F. (2023). Active inductive inference in children and adults: A constructivist perspective. *Cognition, 238*, Article 105471.

Brand, C. O., Mesoudi, A., & Smaldino, P. E. (2021). Analogy as a catalyst for cumulative cultural evolution. *Trends in Cognitive Sciences, 25*(6), 450–461.

Brosnan, S. F., Jones, O. D., Lambeth, S. P., Mareno, M. C., Richardson, A. S., & Schapiro, S. J. (2007). Endowment effects in chimpanzees. *Current Biology, 17*(19), 1704–1707.

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science, 23*(9), 1047–1054.

Carstensen, A., & Frank, M. C. (2021). Do graded representations support abstract thought? *Current Opinion in Behavioral Sciences, 37*, 90–97.

Carstensen, A., Zhang, J., Heyman, G. D., Fu, G., Lee, K., & Walker, C. M. (2019). Context shapes early diversity in abstract thought. *Proceedings of the National Academy of Sciences, 116*(28), 13891–13896.

Christie, S. (2021). Learning sameness: Object and relational similarity across species. *Current Opinion in Behavioral Sciences, 37*, 41–46.

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.

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.

De Lillo, C., Visalberghi, E., & Aversano, M. (1997). The organization of exhaustive searches in a patchy space by capuchin monkeys (Cebus apella). *Journal of Comparative Psychology, 111*(1), 82.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition, 126*(2), 285–300.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition, 130*(3), 335–347.

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.

Dymond, S., & Stewart, I. (2016). Relational and analogical reasoning in comparative cognition. *International Journal of Comparative Psychology, 29*(1), 1–11.

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), Article e22693.

Evans, T. A., Beran, M. J., Harris, E. H., & Rice, D. F. (2009). Quantity judgments of sequentially presented food items by capuchin monkeys (Cebus apella). *Animal Cognition, 12*(1), 97–105.

Fagot, J., & Thompson, R. K. (2011). Generalized relational matching by guinea baboons (Papio papio) in two-by-two-item analogy problems. *Psychological Science, 22*(10), 1304–1309.

Fagot, J., & Tomonaga, M. (1999). Global and local processing in humans (Homo sapiens) and chimpanzees (pan troglodytes): Use of a visual search task with compound stimuli. *Journal of Comparative Psychology, 113*(1), 3.

Felsche, E., Stevens, P., Völter, C. J., Buchsbaum, D., & Seed, A. M. (2023). Evidence for abstract representations in children but not capuchin monkeys. *Cognitive Psychology, 140*, 101530.

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development, 86*(5), 1386–1405.

Flemming, T. M. (2006). *What meaning means for same and different: A comparative study in analogical reasoning*. Thesis: Georgia State University.

Flemming, T. M., Beran, M. J., Thompson, R. K., Kleider, H. M., & Washburn, D. A. (2008). What meaning means for same and different: Analogical reasoning in humans (Homo sapiens), chimpanzees (pan troglodytes), and rhesus monkeys (Macaca mulatta). *Journal of Comparative Psychology, 122*(2), 176.

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.

Forss, S., & Willems, E. (2022). The curious case of great ape curiosity and how it is shaped by sociality. *Ethology, 128*(8), 552–563.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Taylor & Francis Boca Raton.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 47–59.

Gentner, D. (2003). Why we're so smart. In *Language in mind: Advances in the study of language and thought, 195235*.

Gentner, D., Shao, R., Simms, N., & Hespos, S. (2021). Learning same and different relations: Cross-species comparisons. *Current Opinion in Behavioral Sciences, 37*, 84–89.

Girotto, V., Fontanari, L., Gonzalez, M., Vallortigara, G., & Blaye, A. (2016). Young children do not succeed in choice tasks that imply evaluating chances. *Cognition, 152*, 32–39.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature, 410*(6831), 930–933.

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*. Cambridge, MA: Harvard University Press.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages*. Retrieved from http://dippl.org.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science, 24*(2), 87–92.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences, 107*(20), 9066–9071.

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

Haun, D. B., Call, J., Janzen, G., & Levinson, S. C. (2006). Evolutionary psychology of spatial representations in the hominidae. *Current Biology, 16*(17), 1736–1740.

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., Mody, S., & Carey, S. (2016). Infants' representations of same and different in match-and non-match-to-sample. *Cognitive Psychology, 86*, 87–111.

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.

Holyoak, K. J., & Lu, H. (2021). Emergence of relational reasoning. *Current Opinion in Behavioral Sciences, 37*, 118–124.

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.

Hopper, L. M., Kurtycz, L. M., Ross, S. R., & Bonnie, K. E. (2015). Captive chimpanzee foraging in a social setting: A test of problem solving, flexibility, and spatial discounting. *PeerJ, 3*, Article e833.

Hoyos, C., Shao, R., & Gentner, D. (2016). The paradox of relational development: Could language learning be (temporarily) harmful?. In *Paper presented at the CogSci*.

Hutchinson, J. M., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour, 75*(4), 1331–1349.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology, 123*, Article 101334.

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.

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.

Kralik, J. D., & Sampson, W. W. (2012). A fruit in hand is worth many more in the bush: Steep spatial discounting by free-ranging rhesus macaques (Macaca mulatta). *Behavioural Processes, 89*(3), 197–202.

Kroupin, I. G., & 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.

Kroupin, I. G., & Carey, S. E. (2022a). The importance of inference in relational reasoning: Relational matching as a case study. *Journal of Experimental Psychology: General, 151*(1), 224.

Kroupin, I. G., & Carey, S. E. (2022b). You cannot find what you are not looking for: Population differences in relational reasoning are sometimes differences in inductive biases alone. *Cognition, 222*, Article 105007.

Lakshminaryanan, V., Keith Chen, M., & Santos, L. R. (2008). Endowment effect in capuchin monkeys. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 363*(1511), 3837–3844.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version, 1*(1), 3.

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), Article e92160.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology, 21*(2), 269–288.

Nonacs, P. (2001). State dependent behavior and the marginal value theorem. *Behavioral Ecology, 12*(1), 71–83.

Obozova, T., Smirnova, A., Zorina, Z., & Wasserman, E. (2015). Analogical reasoning in amazons. *Animal Cognition, 18*(6), 1363–1371.

Olsson, O., Brown, S., & J.. (2006). The foraging benefits of information and the penalty of ignorance. *Oikos, 112*(2), 260–273.

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.

Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences, 6*(1), 125–136.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development, 13*(4), 453–478.

Ruggeri, A., Swaboda, N., Sim, Z. L., & Gopnik, A. (2019). Shake it baby, but only when needed: Preschoolers adapt their exploratory strategies to the information structure of the task. *Cognition, 193*, Article 104013.

Sánchez-Amaro, A., & Rossano, F. (2023). Comparative curiosity: How do great apes and children deal with uncertainty? *PLoS One, 18*(5), Article e0285946.

Schulz, L., Kushnir, T., & Gopnik, A. (2007). Learning from doing: Intervention and causal inference in children. In A. Gopnik, & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, computation (Ch. 5)* (pp. 67–85). New York, NY: Oxford University Press.

Seed, A., Hanus, D., & Call, J. (2011). Causal knowledge in corvids, primates, and children. *Tool Use and Causal Cognition*, 89–110.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*(4), 325–345.

Sim, Z., & Xu, F. (2015). Toddlers learn with facilitated play, not free play. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2200–2205). Austin, TX: Cognitive Science Society.

Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology, 53*(4), 642–651.

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

Spinozzi, G., De Lillo, C., & Truppa, V. (2003). Global and local processing of hierarchical visual stimuli in tufted capuchin monkeys (Cebus apella). *Journal of Comparative Psychology, 117*(1), 15.

Swait, J., & Marley, A. A. (2013). Probabilistic choice (models) as a result of balancing multiple goals. *Journal of Mathematical Psychology, 57*(1–2), 1–14.

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

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*(6033), 1054–1059.

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.

Tennie, C., Völter, C. J., Vonau, V., Hanus, D., Call, J., & Tomasello, M. (2019). Chimpanzees use observed temporal directionality to learn novel causal relations. *Primates, 60*, 517–524.

Thompson, R. K., & Oden, D. L. (2000). Categorical perception and conceptual judgments by non-human primates: The paleological monkey and the analogical ape. *Cognitive Science, 24*(3), 363–396.

Thompson, R. K., Oden, D. L., & Boysen, S. T. (1997). Language-naive chimpanzees (pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes, 23*(1), 31.

Tomasello, M. (2020). The role of roles in uniquely human cognition and sociality. *Journal for the Theory of Social Behaviour, 50*(1), 2–19.

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

Vonk, J. (2003). Gorilla (Gorilla gorilla gorilla) and orangutan (pongo abelii) understanding of first-and second-order relations. *Animal Cognition, 6*(2), 77–86.

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

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. https://doi.org/10.1177/0956797613502983

Walker, C. M., Rett, A., & Bonawitz, E. (2020). Design drives discovery in causal learning. *Psychological Science, 31*(2), 129–138.

Wang, M. Z., & Hayden, B. Y. (2019). Monkeys are curious about counterfactual outcomes. *Cognition, 189*, 1–10.

Wasserman, E. A., Castro, L., & Fagot, J. (2017). *Relational thinking in animals and humans: From percepts to concepts*.

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.

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.

Zentall, T. R., Andrews, D. M., & Case, J. P. (2018). Sameness may be a natural concept that does not require learning. *Psychological Science, 29*(7), 1185–1189.