Journal of Experimental Child Psychology 200 (2020) 104966

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

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

Problem solving flexibility across early development

Lydia M. Hopper [a,*], Sarah L. Jacobson [b], Lauren H. Howard [c]

[a] Lester E. Fisher Center for the Study and Conservation of Apes, Lincoln Park Zoo, Chicago, IL 60614, USA
[b] Program in Psychology, Graduate Center, City University of New York, New York, NY 10016, USA
[c] Department of Psychology, Franklin & Marshall College, Lancaster, PA 17603, USA

ARTICLE INFO

Article history:
Received 3 April 2020
Revised 24 July 2020
Available online 26 August 2020

Keywords:
Chimpanzee
Comparative cognition
Conservatism
Executive function
Gorilla
Problem solving

ABSTRACT

Cognitive flexibility allows individuals to adapt to novel situations. However, this ability appears to develop slowly over the first few years of life, mediated by task complexity and opacity. We used a physically simple novel task, previously tested with nonhuman primates, to explore the development of flexible problem solving in 2-, 3-, and 4-year-old children from a developmental and comparative perspective. The task goal was to remove barriers (straws) from a clear tube to release a ball. The location of the ball, and therefore the number of straws necessary to retrieve it, varied across two test phases (four of five straws and two of five straws, respectively). In Test Phase 1, all children retrieved the ball in Trial 1 and 83.61% used the most efficient method (removing only straws below the ball). Across Phase 1 trials, 4-year-olds were significantly more efficient than 2-year-olds, and solve latency decreased for all age groups. Test Phase 2 altered the location of the ball, allowing us to explore whether children could flexibly adopt a more efficient solution when their original (now inefficient) solution remained available. In Phase 2, significantly more 4-year-olds than 2-year-olds were efficient; the older children showed greater competency with the task and were more flexible to changing task demands than the younger children. Interestingly, no age group was as flexible in Phase 2 as previously tested nonhuman primates, potentially related to their relatively reduced task exploration in Phase 1.

[*] Corresponding author.
E-mail address: lhopper@lpzoo.org (L.M. Hopper).

https://doi.org/10.1016/j.jecp.2020.104966

Therefore, this causally clear task revealed changes in cognitive flexibility across both early childhood and species.

Introduction

Flexibility allows individuals to nimbly react to novel situations, playing an important role in adaptive responses to environmental changes and finding optimum solutions to problems. Humans are particularly adept at flexible thinking, potentially due to complexity in their environment and social relationship structure (Gökçen, Petrides, Hudry, Frederickson, & Smillie, 2014). Such flexibility is important because it is linked to our innovative ability and tool use (Keen, 2011; Neldner, Mushin, & Nielsen, 2017). Previous research has shown that, as compared with younger children, older children can more flexibly react to environmental or task changes (but see Gopnik et al., 2017). For example, children over 4 years of age can quickly alter sorting techniques on the same objects when given different verbal cues (e.g., "Find the ones that look like a ____" vs. "Find the one that is the same kind as ____"; Deák & Bauer, 1995) or game rules (e.g., "Sort the red ones" vs. "Sort the small ones"; Frye, Zelazo, & Palfai, 1995; Zelazo, Frye, & Rapus, 1996), whereas children under age 4 often perseverate on the first cue or rule with which they are provided, making switch errors on 50–100% of trials (Kirkham & Diamond, 2003). Indeed, children's flexibility appears to develop slowly over the first few years of life and may be mediated by a child's understanding of the task (Karmiloff-Smith, 1990; Spensley & Taylor, 1999). These findings, and others, suggest that cognitive flexibility might be linked to children's biological age (Zelazo, Muller, Frye, & Marcovitch, 2003), although some evidence finds that children as a whole may also be more flexible than adults in certain situations (Lucas, Bridgers, Griffiths, & Gopnik, 2014).

In addition to biological maturation across one's lifespan, we can explore cognitive flexibility from a comparative perspective, for example, by studying nonhuman primates (Sneve et al., 2018). Although there appears to be much variability across and within primate species with regard to their flexible or conservative responses to novel tasks (reviewed in Brosnan & Hopper, 2014), nonhuman primates do exhibit cognitive flexibility in relation to both physical and social understanding (e.g., Amici, Call, Watzek, Brosnan, & Aureli, 2018; Pope et al., 2020). In certain cases, and as compared with adult humans, nonhuman primates (e.g., macaques, capuchin monkeys) have been found to be significantly better at quickly and flexibly altering their behavior in response to changing task demands (e.g., Avdagic, Jensen, Altschul, & Terrace, 2014; Stoet & Snyder, 2007; Watzek, Pope, & Brosnan, 2019). Thus, an exploration both across human development and across species might prove to be particularly insightful for understanding the ontogeny and evolutionary development of cognitive flexibility.

Research with humans and nonhuman primates has revealed that other important factors, including task complexity, opacity, and cognitive demands, influence cognitive flexibility. Zelazo, Carter, Reznick, and Frye (1997) proposed that individuals' ability to evaluate their own success in a task (i.e., error detection and correction) is a key component of problem solving as related to executive function. As such, tasks that make error detection easier are more likely to elicit adequate task switching from children. For example, when 3-year-olds are asked to repeat game rules before task switching, they are much more likely to succeed than when they are asked to simply complete the task without this verbal reminder (Kirkham, Cruess, & Diamond, 2003). Whereas understanding the rules of a task can enhance children's success, causally clear tasks are also more efficiently solved by children and nonhuman primates (Jacobson & Hopper, 2019). For example, 2-year-olds are more successful at an inhibition task if the actions required of them are obviously causal (e.g., pull a lever to get an object) as opposed to unclear or arbitrary (e.g., answer a phone to get an object) (McGuigan & Núñez, 2006). In this way, causal understanding may allow for solutions to be found and errors to be identified as well as flexibility in response to changes in task demands (Hopper, Kurtycz, Ross, & Bonnie, 2015; Jacobson & Hopper, 2019). Research has suggested that differences in response flexibility also likely depend on the cognitive demands inherent in the task. For example, less demanding looking-time paradigms often show much earlier evidence for cognitive flexibility than tasks that require children to act on objects (e.g., Smith, Thelen, Titzer, & McLin, 1999). Given this, (Davis, Schapiro, Lambeth, Wood, & Whiten, 2019) proposed that perseveration is likely mediated by response prepotency (how familiar an action is) and working memory load (how demanding the task is).

Here, we sought to explore how children's cognitive flexibility (i.e., set shifting; Ionescu, 2012) changes across early development and how their strategies compare with those of nonhuman primates (chimpanzees and gorillas tested previously using a comparable task and testing protocol). Specifically, research with nonhuman primates has shown that they are less likely to adopt a novel solution after learning one successful one (Hrubesch, Preuschoft, & van Schaik, 2009) but that this is mediated by task transparency, such that individuals tested with a causally clear task appear to be more flexible (Jacobson & Hopper, 2019). Therefore, unlike many previous studies on early cognitive flexibility, we used a paradigm that was both physically and causally clear and asked children to switch strategies without a large memory demand (e.g., remembering an abstract rule; Zelazo et al., 1996). Specifically, we used a novel puzzle that was a clear vertical tube with five paper straws threaded through it at equal intervals. A small ball was placed in the tube such that it rested on a straw, and to retrieve the ball children needed to pull out all the straws below the ball so that it could fall down the tube and out the bottom. Each time children retrieved the ball, they could exchange it with the researcher for a sticker. Thus, this clear task relied on participants' basic understanding of gravity and support, did not necessitate the use of arbitrary actions, and did not require participants to retain information across trials.

In the first configuration of the task, four straws were below the ball and one was above it. This was to test children's spontaneous understanding of the causal mechanics of the task and to verify that it was causally clear to children. Research shows that even infants appear to have a basic understanding of gravity (Baillargeon & Hanko-Summers, 1990; Needham & Baillargeon, 1993), looking longer at and interacting more with objects if they appear to magically float in space when their support is removed (e.g., Stahl & Feigenson, 2015). Infants also appear to understand that a solid barrier will stop an object from falling or rolling in a downward trajectory even when that object is behind an opaque occluder (Spelke, Breinlinger, Macomber, & Jacobson, 1992). Although 2-year-olds struggle with more cognitively demanding physical adaptation of these looking-time studies, by 3 years of age children are able to track a falling object behind an occluder, correctly select a door to open, *and* reach for the fallen object (Berthier, DeBlois, Poirier, Novak, & Clifton, 2000). Thus, with this first configuration of our task, we could test whether children would remove only straws below the ball and ignore the straw above it or whether they would "blindly" pull out all straws, in turn revealing whether they understood the task rules (without explanation or guidance). Furthermore, by testing 2-, 3-, and 4-year-olds, we could observe whether their understanding and success differed by age.

To examine whether the children could adopt a new solution strategy after repeated experience with the task in the first configuration, we subsequently presented a new configuration where only two straws were below the ball and three straws were above it. In this new configuration, the most efficient solution was to remove two straws instead of four, although the previously efficient strategy remained viable (albeit a less efficient solution). In this way, our paradigm allowed us to test individuals' flexibility in the face of possible conservatism and the interplay between causal understanding and cognitive flexibility across children of different ages and across (primate) species.

Our previous research with nonhuman primates using the same task revealed that chimpanzees and gorillas showed flexible problem solving when task demands changed, likely due to the apes' causal understanding of the task (Jacobson & Hopper, 2019). Therefore, we predicted that if children understood the task mechanics, they would respond flexibly when task demands changed. Specifically, we predicted that if children solved the task using the most efficient strategy in the first task configuration, they would also be able to adopt a new efficient strategy when the configuration was changed. However, we also predicted that the younger children may be less likely to master the task (i.e., understand the solution and so be less likely to use efficient responses) and, accordingly, may be less flexible than the older children (if task understanding relates to flexibility in response patterns). Thus, with our study, we wanted to see how young children responded to changing task demands, how their efficiency and flexibility differed with age, and how their responses compared with those of nonhuman primates.

Method

Participants

We tested 61 children representing three age groups: 20 2-year-olds (M = 30.0 months, range = 24.0–34.7; 10 girls), 22 3-year-olds (M = 40.6 months, range = 36.0–47.8; 12 girls), and 19 4-year-olds (M = 53.4 months, range = 49.0–59.6; 6 girls). From parental reporting, we determined that 81% of participants were Caucasian, 5% were African American, 2% were Hispanic, 2% were Asian American, and 10% were multiracial (2% of parents opted out of answering questions concerning their child's race/ethnicity). In addition to the 61 children described above, we tested 9 children who were not included in the final sample due to refusing to participate in the given tasks (n = 7), failure to obtain video-recording consent (n = 1), or experimenter error (n = 1).

This study received approval from the Franklin & Marshall institutional review board.

Procedure

After entering the testing room, children sat in a chair placed directly in front of Experimenter 1 and the testing apparatus. The apparatus (modeled from the design previously used to test problem solving flexibility in apes; Jacobson & Hopper, 2019) was a clear PVC tube (approximately 2.5 cm in diameter and 63.5 cm long) affixed to a stand with equally spaced holes for up to five straws to be slotted through the tube (Fig. 1). Experimenter 2 sat on a small chair perpendicular to children. Parents sat in a chair opposite from the apparatus, facing children and the experimenters, and were asked to remain quiet and to not intervene in the task or guide children's responses. Sessions were video-recorded for later coding, with the camera located diagonal to the testing area (focused on Experimenter 1, the testing apparatus, and children's hands. Testing comprised a familiarization phase followed by two experimental phases.

In the familiarization phase, Experimenter 1 introduced children to the test apparatus, saying, "This is my toy. Look what this toy can do. When I put a ball in, it comes out the end." The experimenter then proceeded through 3 familiarization trials to acquaint children with the general mechanics of the tube. During these trials, the experimenter dropped a ball into the tube to show how a ball could fall down the apparatus when no obstructions (straws) were present. With each trial, Experimenter 1 verbally prompted children to retrieve the ball that came out of the tube ("Can you get the ball?") and told children that they would be rewarded with a sticker by Experimenter 2 when they obtained the ball ("Every time you give the ball to our friend [Experimenter 2], you get a sticker!"). The sticker provided an incentive for children to quickly retrieve the ball by removing the straw obstructions (in the previous study with chimpanzees and gorillas using this task, a food reward was used in place of the ball here, which was inherently rewarding; see Jacobson & Hopper, 2019).

In Test Phase 1, participants completed 10 trials whereby they were asked to retrieve the ball from the test apparatus. For each trial, Experimenter 1 baited the test apparatus out of view of children (behind a 122 by 91-cm tri-fold display board). Experimenter 1 inserted a straw through each of the five holes in the shaft of the tube, with the ball placed into the apparatus such that there were four straws below the ball and one straw above it (Configuration 1 in Fig. 1). In this configuration, children needed to remove the four straws below the ball to obtain it; the fifth straw could also be removed, but doing so was causally irrelevant to obtaining the ball. The experimenter ensured that the straws were aligned equally to avoid any visual cuing that might encourage participants to select specific straws (e.g., one straw sticking out from the apparatus farther than another straw). Then, the experimenter removed the tri-fold board so that children could view the tube.

A test trial began as soon as the experimenter verbally prompted children to interact with the apparatus ("Can you get the ball?"). Participants were then given a chance to remove any of the straws (below or above the ball) in whatever manner they wished. If children were hesitant to initiate an

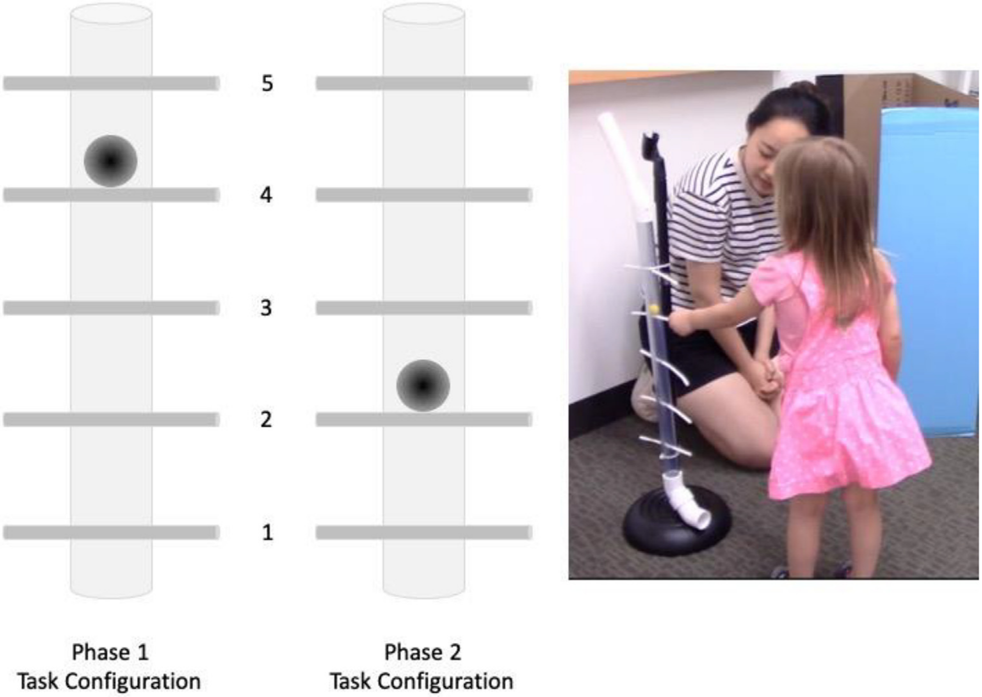

Fig. 1. Schematic of the task—a clear plastic tube with five straws threaded through it—showing the configuration both in Phase 1 (four straws below the ball and one straw above it) and in Phase 2 (two straws below the ball and three straws above it) and a photograph of one participant completing a trial in Phase 1. In either task configuration, participants were free to pull out as many straws us they chose and in any order. Thus, although the most efficient strategy would be to only remove straws below the ball, in either configuration children could also adopt inefficient strategies and remove straws both above and below the ball. Importantly, the efficient solution for Phase 1 (removing Straws 1–4) remained viable in Phase 2, although a different solution was the most efficient one (removing Straws 1 and 2).

interaction with the straws or the apparatus, the experimenter further prompted them with encouraging but noninformative cues (e.g., "It's okay, you can come up and touch it"; "The ball is stuck. How do you think you could get it out?"; "If you get the ball, you'll get a sticker!"). Throughout the trial, the experimenter retained a neutral facial expression and a neutral tone of voice to prevent children from receiving any external cues that would interfere with their interaction with the apparatus. Similar to the familiarization trials, participants received a sticker whenever they acquired the ball and handed it to Experimenter 2. After completing 10 trials with the apparatus in Configuration 1, Phase 2 commenced.

In Test Phase 2, participants completed test trials with the new task configuration. These trials were run as in Phase 1, changing only the apparatus configuration; in Phase 2, the experimenter baited the apparatus such that there were only two straws below the ball and three straws above it (Configuration 2 in Fig. 1). Thus, the total number of straws that *could be* removed (five) was the same across test phases, but the number of straws that *needed to be* removed to obtain the ball differed (four in Phase 1 and two in Phase 2). Importantly, the experimenter never highlighted the change in task configuration, or the new location of the ball at the start of the trial, either verbally or via pointing. As with Phase 1, the experimenter made statements only to encourage children's engagement (e.g., "Can you get the ball?").

In Phase 2, children completed 4 trials. There were two reasons why we ran fewer trials in this second phase. First, our primary interest was assessing children's ability to switch response strategies when the task configuration changed. For this, we were predominantly interested in assessing their responses in the first trial post-configuration change (i.e., to see whether they adopted a new solution and whether they adopted the most efficient solution possible with their first response in Phase 2). Second, although we also wanted to evaluate children's repeated responses in Phase 2 with multiple trials (to see whether their responses increased in efficiency over time if they did not make a strategy switch with their first trial of Phase 2), we did not want to give children too many opportunities to interact with the task because we wanted to test their spontaneous responses. This is in contrast to Phase 1, where we wanted to assess their spontaneous understanding of the task (Trial 1) and also wanted to give them repeated experience with the task across multiple trials both to assess their exploration of the task and to generate a modal response ("remove four straws") that would be more likely to be conserved and potentially harder to deviate from in Phase 2 (in the sense of Jacobson & Hopper, 2019).

Coding

A trained researcher coded all the test trials from video. A second independent research assistant coded 25% of participants' trials, with the two coders agreeing on approximately 99.9% of total behavioral scores. When there was a coding disagreement, we used the primary coder's scoring for a given trial. For each trial, the coder recorded four elements and associated information: the total number of straws participants removed, the order in which participants removed the straws, the length of each trial (i.e., latency to remove straws), and any comments participants made during the first trial of Test Phase 2 when they were presented with a new configuration of the task (i.e., the "switch trial").

Straw removal: Order and number

For each trial, we coded for the total number of straws removed by children (out of a possible five for each trial) and the order in which children removed each straw. If participants retrieved the reward by pulling only the straws below the ball (four straws for Test Phase 1 or two straws for Test Phase 2), we coded the trial as "efficient," but if participants pulled one or more straws above the ball (thereby pulling straws that were not causally necessary to receive the reward), we coded the trial was as "inefficient" (as per Jacobson & Hopper, 2019).

Trial latency

We coded the time at which participants removed each straw within a given trial. Thus, we could calculate the latency for participants to complete each of their trials. The start time for each trial began as soon as the experimenter uttered the introductory prompt ("Can you get the ball?") and ended once children indicated they were done removing straws by explicitly stating such or moving to Experimenter 2 to retrieve a sticker for ball retrieval (typically as soon as the ball fell from the tube).

Verbal responses

During the first trial of Phase 2 (the "switch trial" when the configuration of the apparatus was altered), we transcribed any verbal comments participants made that might indicate that they noticed a change and/or were seeking information related to the change (i.e., "why" questions; Legare, Sobel, & Callanan, 2017). For instance, some participants would recognize the change of the tube arrangement and say, "Why is the ball all the way down there?" or "How did you do that?" We provide a descriptive summary of these in the Results section.

Analysis

To explore participants' spontaneous understanding of the task and their flexibility in response to changing task demands, we analyzed four key aspects of the coded data using R Version 3.5.2 (Core, 2018): (a) participants' spontaneous understanding of the task, (b) any apparent learning across trials, (c) participants' flexibility and efficiency across and within phases, and (d) the verbal responses participants made, if any. For clarity, the specific analytical approaches that we used for each analysis are reported within that section of the Results. For all pairwise comparisons, a Bonferroni correction was applied (i.e., $\alpha \leq .017$). With $\alpha \leq .017$ and an effect size of .80, our power analysis revealed a value of

.63. (To achieve a power value of .80, we would need to have included 29 children per condition, but with the current COVID-19 pandemic, additional testing was not feasible.) We plotted all data using the *ggplot2* package (Wickham, 2016) and *beeswarm* package (Eklund, 2016; see also Wilkinson, 1999) in R Version 3.5.2 (Core, 2018).

Results

Spontaneous understanding of the task

All 61 children tested were able to retrieve the ball from the apparatus in their first trial of Phase 1. Furthermore, 15 of the 20 2-year-olds (75.00%), 18 of the 22 3-year-olds (81.82%), and 18 of the 19 4-year-olds (94.74%) used the most efficient method to do so in Trial 1 of Phase 1 (i.e., they removed only the lower four of the five straws from the tube) (Fig. 1). We compared children's spontaneous ability to solve the task across the three age groups. To do so, we used first trial efficiency (efficient or inefficient) as our outcome variable and participant ID as a random factor. Given the binary response variable, we analyzed our data using a binomial generalized linear mixed model (GLMM) in R Version 3.5.2 (Core, 2018). We fit this model using the Laplace approximation method via the "glmer" function in the *lme4* package (Bates, Maechler, & Bolker, 2012) to test the relative effect of our predictor variable age group (family = "binomial"). This revealed that there was no significant difference across the three age groups of children in their likelihood to use the most efficient method in their first trial of Phase 1 {z = 0.34, p = .733, 95% confidence interval (CI) [−3.80, 5.41]}. Not only were the three age groups equally likely to use an efficient response with their first trial, but of those children who responded efficiently in the first trial of Phase 1, there was strong consistency in the action sequence (i.e., straw removal order) that they used: 93.33% of the 2-year-olds who responded efficiently used the 4,3,2,1 action sequence, as did 100.00% of the 3-year-olds and 94.44% of the 4-year-olds who responded efficiently (i.e., they sequentially removed the straw directly below the ball) (Fig. 1).

Proficiency with the task in Phase 1

On average across all their trials in Phase 1, 2-year-olds used an efficient action sequence in 60.56% (SD = 27.75) of trials, whereas 3-year-olds were efficient in 71.90% (SD = 32.19) of trials and 4-year-olds were efficient in 86.32% (SD = 24.99) of trials (Fig. 2). Using a GLMM, we explored the proportion of all the participants' trials in Phase 1 that were efficient (family = "poisson") and used independent t tests, using the "t.test" function to compare children's efficiency across age groups. This revealed that there was a significant effect of age on the percentage of trials in which children made efficient responses in Phase 1 (z = 2.80, p = .005, 95% CI [0.53, 0.30]). Specifically, 4-year-olds made significantly more efficient responses than 2-year-olds, t(34.13) = −2.96, p = .006, 95% CI [−4.34, −0.81], but there was no significant difference in the percentages of trials in which 2- and 3-year-olds made efficient responses, t(36.99) = −1.18, p = .245, 95% CI [−3.08, 0.81] or in the percentages of trials in which 3- and 4-year-olds made efficient responses, t(37.19) = −1.59, p = .120, 95% CI [−3.28, 8.63].

When first presented with the task in Phase 1, the average latency for 4-year-olds to complete their first trial was 23.05 s (SD = 10.43). The 4-year-olds were significantly quicker to complete their first trial than the 2-year-olds (average latency = 80.35 s, SD = 71.98), t(19.84) = 3.52, p = .002, 95% CI [23.34, 91.26] and the 3-year-olds (average latency = 61.95 s, SD = 70.03), t(22.08) = 2.57, p = .017, 95% CI [7.55, 91.26]. There was no significant difference, however, between the 2- and 3-year-olds in the time it took them to complete their first trial, t(39.38) = 0.84, p = .407, 95% CI [−26.00, 62.79]. To test whether children's trial completion times became quicker across trials, as a proxy for learning, we correlated participants' trial latency with trial number using the "rmcorr" function (Bakdash & Marusich, 2017). This takes into account repeated samples from participants to determine whether their trial latency decreased over time. This revealed that there was a significant negative correlation between the trial completion latency and trial number for all three age groups, such that children became quicker to complete trials across the 10 trials in Phase 1: 2-year-olds (r = −.385, p < .001), 3-year-olds (r = −.297, p < .001), and 4-year-olds (r = −.232, p = .002). The weaker negative

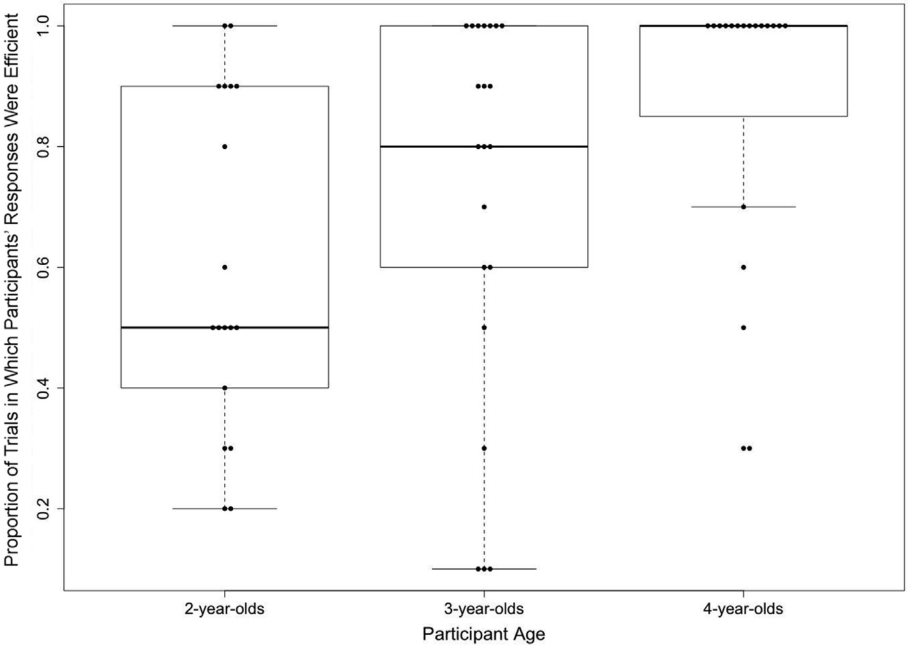

Fig. 2. Proportions of Phase 1 trials in which the three age groups of children used an efficient solution.

relationship between trial number and latency for 4-year-olds is likely because they completed their first trial faster than the younger children and there is likely a limit to how quickly any children can complete a trial, creating a floor effect. Indeed, 4-year-olds' average trial completion latency delta from Trial 1 to Trial 10 was only 14.89 s (average Trial 10 latency = 8.16 s, *SD* = 23.05), whereas the delta for 3-year-olds was 47.90 s (average Trial 10 latency = 14.05 s, *SD* = 15.51) and for 2-year-olds was 64.35 s (average Trial 10 latency = 16.00 s, *SD* = 10.83).

Strategy-switching flexibility and efficiency in response to changed task configuration

To evaluate children's cognitive flexibility, we assessed their efficiency in the first trial of Phase 2 when they were presented with the new task configuration (Fig. 1). Only 7 (11.48%) of the children used the same action sequence (straw removal order) in the first trial of Phase 2 as they had used in their last trial of Phase 1. Specifically, 4 2-year-olds and 2 3-year-olds used the 5,4,3,2,1 action sequence in both trials, whereas 1 4-year-old used the 4,3,2,1 action sequence in the last trial of Phase 1 and the first trial of Phase 2. Thus, the majority of children (88.52%) used a different action sequence across these trials. For all children and action sequences used, in the first trial of Phase 2, 7 of the 20 2-year-olds (35.00%), 14 of the 22 3-year-olds (63.64%), and 13 of the 19 4-year-olds (68.42%) used the (newly available) most efficient method (i.e., they removed only the lower two of five straws from the tube), highlighting their recognition of the changed task demands. As with Test Phase 1, we used the "glmer" function in the *lme4* package (family = "binomial") to compare the numbers of children across the three age groups whose first trial in Phase 2 was efficient, and we used independent *t* tests using the "t.test" function for post hoc pairwise comparisons across age groups. Our analyses revealed that there was a significant effect of age on children's efficiency in the first trial of Phase 2 (z = 2.74, p = .023, 95% CI [0.11, 1.51]). In spite of this, after correcting for multiple comparisons, post hoc pairwise comparisons revealed no significant difference across age groups when comparing the numbers of children whose responses in the first trial of Phase 2 responses were efficient: 4-year-olds versus 2-year-olds, t(35.92) = −2.41, p = .021, 95% CI [−0.68, −0.06]; 4-year-olds versus 3-year-olds,

$t(37.27) = 0.57$, $p = .573$, 95% CI [−0.22, 0.39]; 2-year-olds versus 3-year-olds, $t(39.67) = -1.89$, $p = .066$, 95% CI [−0.59, 0.02].

Considering all 4 trials that children completed in Phase 2, on average children removed significantly fewer straws per trial in Phase 2 than they did in Phase 1, highlighting their understanding of the changed task demands. This was true for all three age groups of children tested: 2-year-olds, $t(21.45) = 3.81$, $p = .001$, 95% CI [0.49, 1.68]; 3-year-olds, $t(25.10) = 6.20$, $p < .001$, 95% CI [1.04, 2.08]; 4-year-olds, $t(20.65) = 7.55$, $p < .001$, 95% CI [1.21, 2.13]. Although children removed *fewer* straws in Phase 2 as compared with Phase 1, did they consistently remove the *fewest* possible number (i.e., two straws)? On average across all trials in Phase 2, 2-year-olds used an efficient action sequence in 42.50% ($SD = 39.82$) of their trials, whereas 3-year-olds used an efficient solution in 68.18% ($SD = 41.68$) of trials and 4-year-olds were efficient in 81.58% ($SD = 32.10$) of trials. Using a GLMM (family = "poisson"), we found that there was a significant effect of age on the proportion of Phase 2 trials in which children made efficient responses ($z = 3.04$, $p = .002$, 95% CI [0.11, 0.51]). Specifically, 4-year-olds made significantly more efficient responses than 2-year-olds, $t(36.07) = -3.38$, $p = .002$, 95% CI [−2.50, −0.63], but there was no significant difference in the proportions of trials in which 3- and 4-year-olds made efficient responses, $t(38.54) = -1.16$, $p = .253$, 95% CI [−1.47, 0.40], or in the proportions of trials in which 3- and 2-year-olds made efficient responses, $t(39.89) = -2.04$, $p = .048$, 95% CI [−2.04, −0.01].

To further explore children's causal understanding of the task and their ability to flexibly shift strategies across the phases in response to the change in task configuration, we compared children's latency to complete trials across the two phases as a proxy for flexibility (i.e., removing two straws should take less time than removing four straws). Across all children tested, they were significantly faster to complete trials in Phase 2 (average trial completion latency = 12.03 s, $SD = 13.36$) compared with Phase 1 (average trial completion latency = 17.44 s, $SD = 11.45$), $r(60) = -.426$, $p < .001$. There was also a significant effect of age on children's latency to complete a trial. Within Phase 2, the average latency for 2-year-olds to complete a trial was 17.40 s ($SD = 16.70$), whereas the average trial completion latencies for 3- and 4-year-olds were 11.82 s ($SD = 13.63$) and 6.63 s ($SD = 4.61$), respectively. The 4-year-olds completed trials significantly faster than both the 2-year-olds, $t(49.82) = 4.74$, $p < .001$, 95% CI [7.10, 17.57] and 3-year-olds, $t(63.56) = 3.31$, $p = .002$, 95% CI [2.65, 10.68]. In contrast, there was no significant difference between the 2-year-olds' and 3-year-olds' trial completion latency, $t(73.10) = 1.88$, $p = .056$, 95% CI [−0.35, 11.69].

In addition to comparing children's understanding of the task and flexibility across ages, we were also interested in how consistently proficient each child was. To examine this, we compared children's efficiency in Phase 1 with their efficiency in Phase 2. We found that, for all three age groups, children showed intra-individual consistency in their efficiency across phases; that is, the proportion of trials that children solved efficiently in Phase 1 was significantly correlated with the proportion of trials that children solved efficiently in Phase 2: Pearson's product–moment correlation, 2-year-olds, $t(18) = 5.066$, $p < .001$, 95% CI [.49, .90]; 3-year-olds, $t(20) = 3.635$, $p = .002$, 95% CI [.28, .83]; 4-year-olds, $t(17) = 6.093$, $p < .001$, 95% CI [.60, .93].

Conservatism and diversity of solution strategies

As reflected by children's responses in the first trial of Phase 1, the action sequence most commonly used by children across all trials in Phase 1 was repeatedly removing the straw directly below the reward (i.e., 4,3,2,1) (Figs. 1 and 3). This action sequence represented 53.78% of 2-year-olds' trials, 64.84% of 3-year-olds' trials, and 73.82% of 4-year-olds' trials in Phase 1. In addition, and as can be seen in Fig. 3, the modal inefficient action sequence for all three age groups in Phase 1 was 5,4,3,2,1 (i.e., pulling out all the straws from top to bottom).

Not only was there consistency across children in their modal action sequence (4,3,2,1) in Phase 1, there was also intra-individual consistency such that some children perseverated in their response phenotype and used an action sequence in multiple successive trials. Indeed, 14 children (1 2-year-old, 5 3-year-olds, and 8 4-year-olds) used the same efficient action sequence for every response they made in Phase 1, and 2 2-year-olds used the same inefficient sequence in each of their 10 trials. Therefore, we explored children's conservatism in this regard. For each child, we calculated the longest run

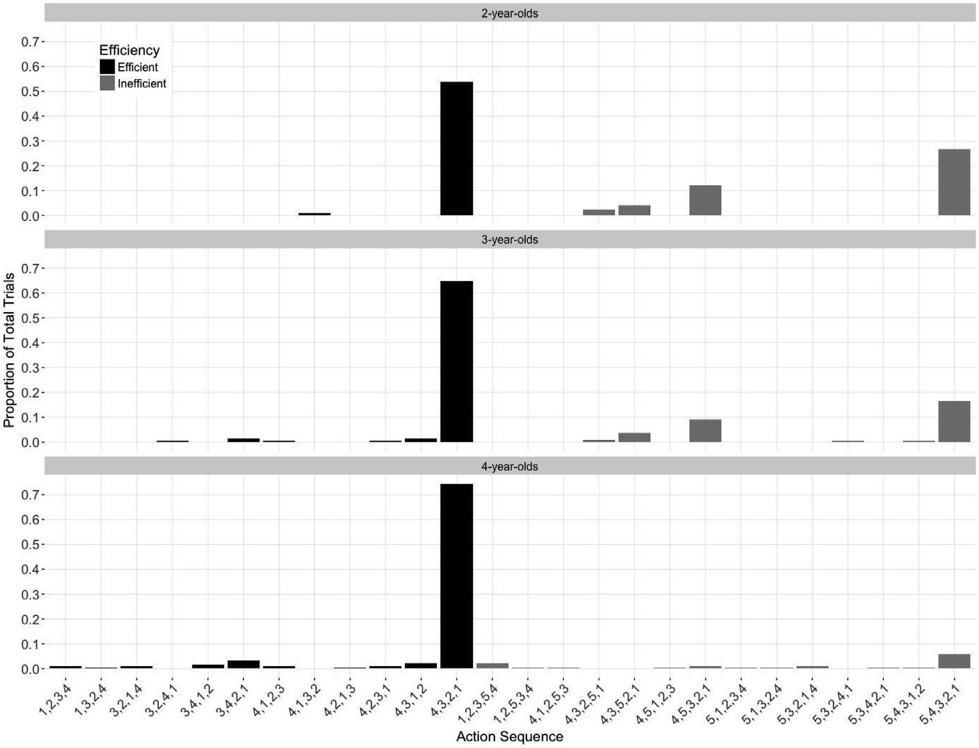

Fig. 3. Proportions of trials in which each action sequence type (i.e., straw removal order)—efficient and inefficient—was used by the 2-, 3-, and 4-year-old children in Phase 1.

of consecutive trials in which the child used the same action sequence in Phase 1, where 2 consecutive trials was the shortest possible run length and 10 consecutive trials was the longest possible run length. In addition, we coded whether each run was efficient or inefficient, and we calculated each child's average efficient run length as a proportion of total possible responses. From this, we found that 4-year-olds made significantly longer efficient runs (average proportion of trials = .63, *SD* = .41) than 2-year-olds (average = .30, *SD* = .27), *t*(25.32) = −2.752, *p* = .011, 95% CI [−.57, −.08] (Fig. 4). However, there was no significant difference in the average efficient run length made by 4- and 3-year-olds (average = .53, *SD* = .36), *t*(30.03) = − 0.778, *p* = .443, 95% CI [−.36, .16] or in the average efficient run length made by 2- and 3-year-olds, *t*(38.31) = −2.315, *p* = .026, 95% CI [−.43, −.03].

In spite of the aforementioned conservatism shown by some children to perseverate on an action sequence across multiple trials, there was variation in the action sequences used by the children across trials. Thus, even after discovering the 4,3,2,1 solution, children sampled other action sequences. Collectively, children used 27 different action sequences in Phase 1 (120 were possible) (Fig. 3). They used 12 (50.00%) of the 24 possible action sequences that were efficient, removing Straws 1, 2, 3, and 4 first (e.g., 1,3,2,4 and 4,3,2,1), but used only 15 (15.63%) of the 96 possible action sequences that were inefficient in which they pulled out the irrelevant straw before releasing the reward (e.g., 5,4,3,2,1 and 3,5,4,2,1). In addition, 2 (33.33%) of the 6 different action sequences that 2-year-olds used were efficient ones, whereas 6 (50.00%) of the 12 action sequences used by 3-year-olds and 10 (45.45%) of the 22 action sequences used by 4-year-olds were efficient.

To determine the diversity of action sequences children used, we calculated the diversity index of their responses (Shannon & Weaver, 1949). If participants repeatedly used the same action sequence (i.e., developed a habit), their diversity index would be lower than those who did not. We used

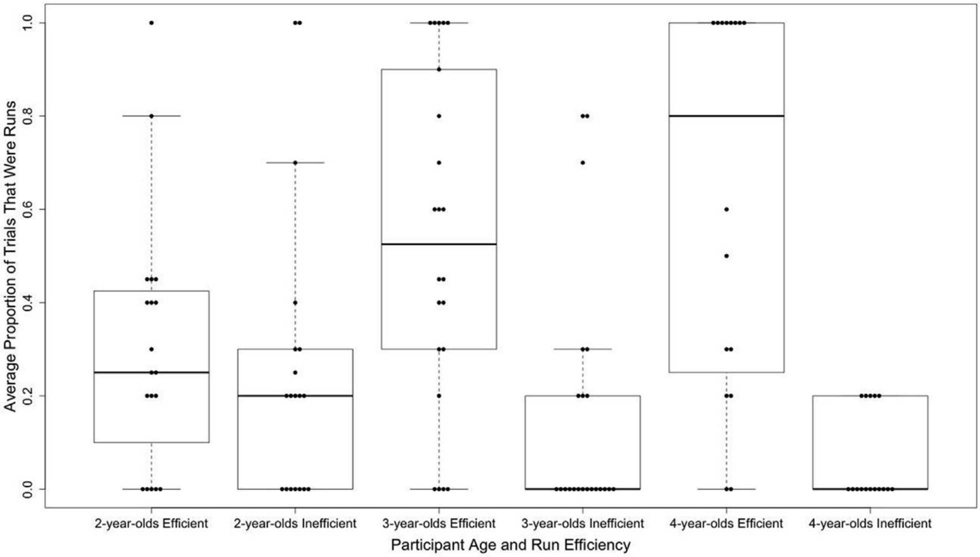

Fig. 4. Average run lengths in Phase 1, as proportions of total trials, for each of the three age groups of children showing both runs comprising efficient and inefficient action sequences.

Wilcoxon tests ("wilcox.test") to compare participants' "H-index" diversity index across age groups and across test phases. Children's individual H-index scores in Phase 1 ranged from 0.00 to 2.03, where an index score of 0 means that only one action sequence was used and an index score of 2.30 would mean that a different action sequence was used for each of the 10 trials, although this never occurred, as indicated by children's maximum score of 2.03. In Phase 1, there was no significant difference in children's H-index scores across the three age groups: 2-year-olds versus 3-year-olds, $t(39.12) = 0.443$, $p = .661$, 95% CI [−.22, .34]; 3-year-olds versus 4-year-olds, $t(27.40) = -0.258$, $p = .799$, 95% CI [−.46, .36]; 2-year-olds versus 4-year-olds, $t(28.93) = 0.047$, $p = .963$, 95% CI [−.41, .43]. In spite of this, we found differences across the three age groups in the relationship between their Phase 1H-index score and the proportion of trials in which they used an efficient action sequence. Specifically, for 4-year-olds, there was a significant negative correlation between their H-index score and their proportion of trials that were efficient {Pearson's product–moment correlation: $t(17) = -6.469$, $p < .001$, 95% CI [−.94, −.63]}, and this was also the case for 3-year-olds, $t(20) = -3.373$, $p = .003$, 95% CI [−.82, −.24], but not 2-year-olds, $t(18) = -1.137$, $p = .271$, 95% CI [−.63, .21].

In contrast to Phase 1, children needed to complete only 4 trials in Phase 2, so there was a higher probability that they would use the same response in all trials. Indeed, whereas only 16 children (26.23%) used the same response across all trials in Phase 1 (described above), 29 children (47.54%) used the same response across all trials in Phase 2, and 24 of these children used an efficient response for every trial (no 4-year-olds used the same inefficient action sequence repeatedly across trials). Reflecting this intra-individual consistency, there was also inter-individual consistency in the specific action sequence that children used in Phase 2. As in Phase 1, the action sequence most commonly used by children in Phase 2 was repeatedly removing the straw directly below the reward (i.e., 2,1) (Fig. 5). This action sequence represented 41.25% of 2-year-olds' trials, 62.50% of 3-year-olds' trials, and 75.00% of 4-year-olds' trials, and as in Phase 1 the modal inefficient response for 2- and 3-year-olds was to remove all the straws from top to bottom (i.e., 5,4,3,2,1) (Fig. 5).

In spite of children's preference for the 2,1 action sequence in Phase 2, they still explored other solution phenotypes, including the alternative efficient action sequence (1,2) and an additional 25 different inefficient action sequences (Fig. 5). In addition to the 2 efficient strategies, 2-year-olds used 17 inefficient action sequences (22.22% of their action sequences were efficient), whereas 3-year-olds

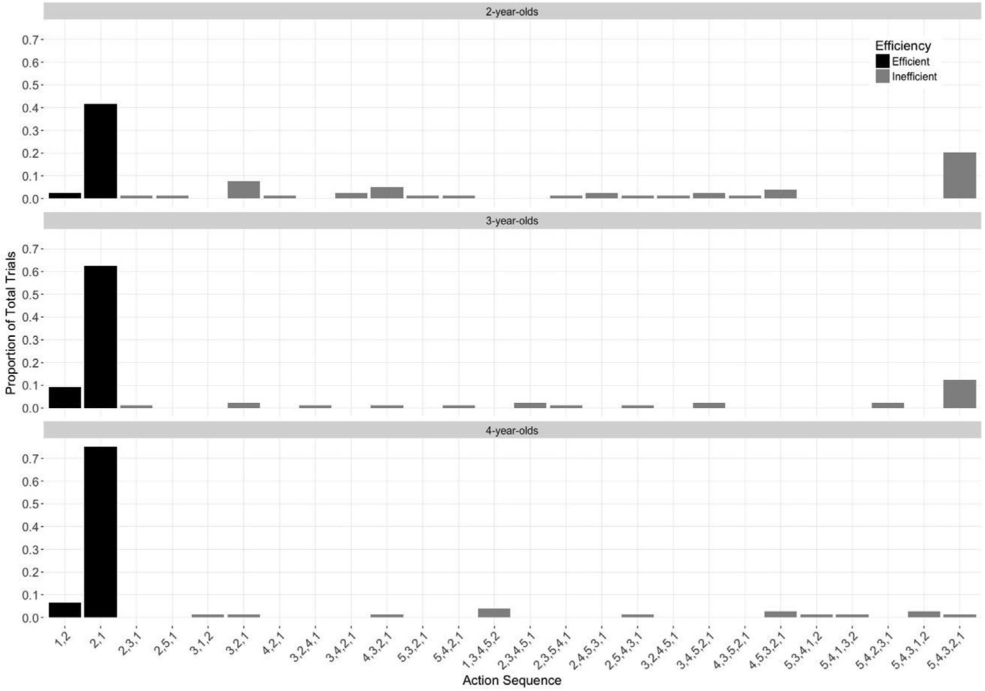

Fig. 5. Proportions of trials in which each action sequence type (i.e., straw removal order)—efficient and inefficient—was used by the 2-, 3-, and 4-year-old children in Phase 2.

used 11 inefficient sequences and 4-year-olds used 10, meaning that 15.38% and 16.67% of the action sequences they used were efficient, respectively. Children's individual H-index scores in Phase 2 ranged from 0.00 to 1.39, where an index score of 0 means that only 1 action sequence was used and an index score of 1.39 would mean that a different action sequence was used for each of the 4 trials. As in Phase 1, in Phase 2 there was no significant difference in children's H-index scores across the three age groups: 2-year-olds versus 3-year-olds, $t(37.42) = 1.768$, $p = .085$, 95% CI [−.04, .60]; 3-year-olds versus 4-year-olds, $t(39.00) = 0.089$, $p = .930$, 95% CI [−.26, .29]; 2-year-olds versus 4-year-olds, $t(34.79) = 1.894$, $p = .067$, 95% CI [−.02, .61]. However, as with their responses in Phase 1, we found that the older children's diversity of responses (H-index score) was negatively correlated with efficiency. Specifically, for 4- and 3-year-olds, there was a significant negative correlation between their H-index score and their proportion of trials in Phase 2 that were efficient {4-year-olds: $t(17) = -3.101$, $p = .006$, 95% CI [−.83, −.20]; 3-year-olds: $t(20) = -2.623$, $p = .016$, 95% CI [−.76, −.11]}, but this was not the case for 2-year-olds, $t(18) = -1.776$, $p = .093$, 95% CI [−.71, .07].

Verbal responses

During the first trial of Phase 2, when children were first presented with the novel configuration of the task (i.e., the "switch trial"), none of the 2-year-olds made any verbal comment in relation to the task. However, 36.36% of the 3-year-olds did, as did 47.37% of the 4-year-olds. Most 3-year-olds' comments reflected the change in task configuration (e.g., "Why did it go down to this one?"; "It's not up there anymore"), whereas other comments highlighted the configuration change but also flagged the experimenter's agency in causing that change (e.g., "Why did you put two?"; "How did you do that?"). Like 3-year-olds, 4-year-olds' comments referred to the change of task configuration (e.g., "There's only two straws"; "Hey, it's now down there") and the experimenter's causation of the change (e.g., "How did you do that?"), but 4-year-olds also commented on how this change affected their own behavior and success (e.g., "I only had to get two"; "That was super fast—that's because there were only two straws"). Both 3- and 4-year-olds commented on the configuration change in 46.67% of their first trials in Phase 2 in which they made an efficient response (i.e., removing only the bottom two straws). In the trials in which children made an inefficient response (i.e., removing three or more straws), 14.29% of 3-year-olds commented on the change, whereas a quarter (25.00%) of 4-year-olds commented on the change. For both age groups, there was no significant difference in the numbers of efficient first trial responses in which children made a comment on the task configuration (Fisher's exact test: 3-year-olds, p = .193; 4-year-olds, p = .399).

Discussion

In our study, we explored 2-, 3-, and 4-year-old children's ability to flexibly switch between response patterns as task demands changed. As Jacobson and Hopper (2019) found previously for nonhuman primates, all the children easily mastered the task and retrieved the ball from the tube. However, 4-year-olds were consistently more efficient than the younger children in terms of both the time it took them to complete trials and the number of straws they removed. This developmental trajectory in children's responses reflects previous research showing that children's problem solving and tool making skills increase with age (Gönül, Takmaz, Hohenberger, & Corballis, 2018). Indeed, the cognitive complexity and control theory proposes that "executive function can be understood in terms of age-related increases in the maximum complexity of the rules children can formulate and use when solving problems" (Zelazo et al., 2003, p. 274). In spite of this, there was no significant difference across the three age groups of children in their likelihood of using the newly available efficient solution when it was presented in the first trial of Phase 2. We also identified intra-individual consistency in children's success such that their efficiency in Phase 1 correlated with their efficiency in Phase 2. Given the general success of children in all age groups, it is likely that this causally clear task facilitated children's success and flexibility (in the sense of Davis et al., 2019), as has been found in chimpanzees and gorillas tested using the same task (Jacobson & Hopper, 2019).

Our aim was to provide a task that was accessible for children in all three age groups to allow us to make meaningful across-age comparisons (as well as comparisons with nonhuman primates' responses). Supporting our goal, all children spontaneously solved the task and there was no difference across the three age groups in their initial understanding of the task, as evidenced by their comparable likelihoods to use the most efficient method in their first trial of Phase 1 (83.61% of children used an efficient solution in the first trial of Phase 1). However, 4-year-olds showed sustained efficiency across trials in Phase 1; significantly more of their trials were solved via the efficient method than those of 2-year-olds. Although the 4-year-olds were not more likely to spontaneously use the most efficient solution in their first trial of Phase 1 than the younger children, they were more likely to stick with it and were significantly quicker to complete their first trial, suggesting enhanced physical dexterity, potentially in combination with a better understanding of the task demands. However, and in spite of 4-year-olds' greater use of efficient solutions, within Phase 1 all three age groups showed an improvement in task proficiency over time, as demonstrated by the negative correlation between trial latency and trial number.

Although the vast majority of children spontaneously used an efficient solution when first presented with the task, when we changed the task demands and introduced the possibility of a new efficient solution in Phase 2, only 57.73% of children spontaneously used the most efficient solution with their first attempt (removing only two straws). As with the first trial of Phase 1, however, there was no effect of age on children's likelihood to use an efficient solution for the first trial of Phase 2. Thus, nearly half of the children, irrespective of age, did not switch strategies in the first trial of Phase 2. In certain situations, humans react in remarkably fixed ways even when their environment does not necessitate such rigidity (Bilalić, McLeod, & Gobet, 2008; Gopnik, Griffiths, & Lucas, 2015). For example, adults often sit in the same seat during classes or meetings even without seat assignment and when there are no repercussions for moving (Costa, 2012). The limited flexibility we observed cannot be explained by pure conservatism (in the sense of Hrubesch et al., 2009); of the 26 children who used an inefficient solution for the first trial of Phase 2, only 7 (26.92%) were inefficient because they used the *exact same* action sequence as they had used in the previous trial (i.e., the final trial of Phase 1). We also note that a large subset of 3- and 4-year-olds remarked on the change in task configuration. Although children often verbally seek out information to understand causal elements of their environment (Legare et al., 2017), our experimenters were instructed not to answer children's questions and were not useful as informants (and, anecdotally, children almost never explicitly asked for help). Therefore, it is possible that children were describing the changes they observed to help themselves make sense of the changes and respond to the new task demands (Winsler, Fernyhough, & Montero, 2009).

We had predicted that greater exploration of the task (i.e., using a range of action sequences) would protect children against conservatism and allow them to more flexibly respond when task demands changed. Paradoxically, although 4-year-olds were the most efficient in their responses, they were also the most conservative; fully 73.82% of their trials in Phase 1 were solved using the same action sequence (4,3,2,1), and we saw similar patterns in their responses in Phase 2 (when they preferentially used the 2,1 action sequence). Indeed, the older children's preferred responses (whether efficient or inefficient) were to remove straws sequentially rather than in a random pattern (although they did this on occasion). Reflecting this, of the 14 children who used the same action sequence for every trial in Phase 1, 8 were 4-year-olds (only 1 2-year-old used the same action sequence in every trial in Phase 1). Indeed, for 3- and 4-year-olds, but not 2-year-olds, there was a significant negative correlation between their H-index score and proportion of trials that were efficient. Thus, for the older children, decreased diversity was associated with increased efficiency, reflecting the results of their likelihood to display longer runs of efficient action sequences (this was also seen in their Phase 2 responses). In this way, the younger children's greater exploration did not benefit them either within phases or in their flexibility across phases. Although conservatism is often seen as a sign of reduced cognitive flexibility, because the older children *struck on* and then *stuck with* an efficient solution from the start, they were able to sustain their efficiency (i.e., "If it ain't broke, don't fix it"). Furthermore, the older children's apparent flexibility in switching strategies from Phase 1 to Phase 2 might not reflect a switch but rather a continuation of the same strategy ("Pull the straw below the ball"). Without further controls, it is difficult to discern whether this is an insightful solution or a rigid response—sticking with the first reinforced pattern used—but we would argue the former given that the younger children also used this solution early on but did not stick with it.

The increased exploration in the younger children aligns with some of the ideas outlined in the *overlapping waves theory*, which states that children do not simply progress from ignorance to full comprehension across development but rather proceed through cognitive waves involving data collection, mapping, strengthening, and refinement when attempting to effectively solve problems (e.g., Chen, Siegler, & Daehler, 2000). However, although this exploration might be viewed favorably with respect to cognitive flexibility, it is actually counter to maximizing efficiency. In Phase 2, there were fewer possible action sequences that were efficient as compared with Phase 1, and so we might expect that children with a strong causal understanding of the task, and a desire to be efficient, would use fewer different action sequences in Phase 2 than in Phase 1. Indeed, this is what we saw with 4-year-olds. The 4-year-olds used fewer action sequences in Phase 2 (12) than in Phase 1 (22), whereas the 3-year-olds used a comparable number of action sequences in both phases (12 vs. 13). Although our a priori goal for this task was to remove the ball by removing as few straws as possible, we never explicitly shared this goal with the children. Therefore, for the younger children, rather than maximizing efficiency, play and exploration might have been stronger drives, which can be advantageous (Greve & Thomsen, 2016). A drive to play might explain why the younger children used more action sequences in Phase 2, although this could also be explained by a reduced understanding of the task mechanics or could be related to young children's tendencies to be more exploratory when events are surprising (Stahl & Feigenson, 2017).

In addition to exploring ontogenetic changes in children's cognitive flexibility, we were also interested in comparing children's behavior in this task with that of chimpanzees and gorillas tested previously with the same task under comparable protocols. When first presented with the task, all children spontaneously retrieved the ball and 83.61% used the most efficient method with their first attempt. As noted above, the apes we tested previously were equally successful, with 84.62% of them using the most efficient method when first presented with the same task (Jacobson & Hopper, 2019). However, in spite of the seeming similarities across species, there were differences in the *way* in which the children and apes solved the task. For example, whereas the children used 12 different efficient action sequences collectively in Phase 1, the apes deployed 21. Furthermore, the children used fewer action sequences on average across the first 10 trials of Phase 1 as compared with the apes (see Fig. S1 in the online supplementary material). The increased exploration by the apes may be due to differences in experimental protocol (children completed all 10 trials within a single session, whereas apes completed trials over one or more sessions; see Jacobson & Hopper, 2019, for details). However, it is notable that whereas the 2- and 3-year-old children never used more than 4 different action sequences each, the 4-year-old children used up to seven and eight different sequences each, revealing exploration rates more similar to that of the apes.

A greater percentage of the apes were flexible in adopting a more efficient response in the first trial of Phase 2 as compared with the children even when comparing apes with the oldest child age group. As discussed, within Phase 1 and across all three age groups, the children predominantly solved the task by repeatedly removing the straw directly below the ball (i.e., 4,3,2,1). This was also the predominant strategy used by the chimpanzees tested previously, but not by the gorillas, whose preferred strategy was to remove straws sequentially from the bottom up (i.e., 1,2,3,4) (cf. Fig. 3 here with Fig. 2 in Jacobson & Hopper, 2019). The strategy of consistently moving the straw that the ball rests on could potentially represent a simple association that was learned by the children rather than a holistic understanding of the task mechanism, but what explains these apparent species differences is not clear at this time. The observed species differences may be a result of methodological elements between this study and that of Jacobson and Hopper (2019). Namely, we gave the children 10 trials in Phase 1 before changing the task configuration, whereas the apes received more than 20 trials spread over multiple sessions before the task was changed. The apes' increased experience with the task may have afforded them greater experience with the task, which may have allowed them to be more flexible or simply gave them more opportunities to explore alternative action sequences. Indeed, this might be why the apes' average run length was shorter than that of the children (average proportion of apes' first 10 trials that were runs = .21, *SD* = .13; cf. with Fig. 4 here); however, unlike the children, the apes never performed a run of inefficient action sequences in their first 10 trials of Phase 1 (see Table 2 in Jacobson & Hopper, 2019). Future work is needed to tease apart the influences of experience, causal understanding, and conservatism on the apparent species differences we identified.

From our results, we propose that causal understanding of a task not only promotes problem solving but also reduces the likelihood of conservative perseveration (see also Jacobson & Hopper, 2019). However, humans are inherently social, and although we may sometimes solve problems by ourselves, we also often seek out information from others. In a landmark study, Bonawitz et al. (2011) found that children were much less flexible when they were directly trained on how to use a certain object. Termed the "double-edged sword of pedagogy," children who observed someone interacting with the object were more likely to explore and learn its multiple functions, whereas those who were directly given instructions did not stray from the singular function they were taught. This effect has now been seen in a number of other contexts, such as children learning to flexibly solve new math problems (Loehr, Fyfe, & Rittle-Johnson, 2014), and may explain children's proclivity for overimitation (Lyons, Young, & Keil, 2007; Over & Carpenter, 2013; Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009). Thus, although direct social instruction can help children to quickly learn how to complete a task, it might also unnecessarily cause behavioral perseveration. Conversely, nonhuman primates appear to be less influenced by social norms as compared with children (e.g., Haun, Rekers, & Tomasello, 2014; Horner & Whiten, 2005; but see Hopper, Schapiro, Lambeth, & Brosnan, 2011). Future research could explore the role of individuals' causal understanding and their reliance on social information on cognitive flexibility (e.g., Burdett, McGuigan, Harrison, & Whiten, 2018) from both a comparative perspective and an ontogenetic perspective (e.g., Horner & Whiten, 2005; Pope, Fagot, Meguerditchian, Washburn, & Hopkins, 2019; Stengelin, Hepach, & Haun, 2020; Wood, Kendal, & Flynn, 2013).

Here, we found that although all three age groups of children were successful in solving the task, 4-year-olds were more efficient and more flexible in their approach to solving the task and responding to changing task demands than 2-year-olds. We previously tested apes on the same task and concluded that their ability to alter the solution strategy they used when we changed the task configuration was likely linked to their causal understanding of the task (Jacobson & Hopper, 2019). Unfortunately, procedural differences in testing protocols across species prevents us from making too many inferences about the apparent species differences we observed or what might drive these differences. However, we note that research using different tests of cognitive flexibility has also identified differences across human and nonhuman primates' responses to matched tasks (e.g., Avdagic et al., 2014; Pope et al., 2020; Watzek et al., 2019), although typically such research has tested adult humans, not young children as we did. Future work exploring the interplay among social information, causal understanding, and cognitive flexibility is needed.

CRediT authorship contribution statement

Lydia M. Hopper: Conceptualization, Methodology, Formal analysis, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Sarah L. Jacobson:** Conceptualization, Methodology, Writing - review & editing. **Lauren H. Howard:** Methodology, Investigation, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

A portion of this study was funded by the Franklin & Marshall College Hackman Scholars Program award to LHH, and LMH received support from the Lincoln Park Zoo's Women's Board. In addition, we thank the following individuals for their assistance with data collection, coding, and methods preparation: Lauren Hein, Natalie Hutchins, Enya Meade, You Jin Park, Ayla Saferstein, Tian Tian, Ellen Verry, Fiona Waters, and Peiru Yu.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2020.104966.

References

Amici, F., Call, J., Watzek, J., Brosnan, S., & Aureli, F. (2018). Social inhibition and behavioural flexibility when the context changes: A comparison across six primate species. *Scientific Reports, 8*, 3067.

Avdagic, E., Jensen, G., Altschul, D., & Terrace, H. S. (2014). Rapid cognitive flexibility of rhesus macaques performing psychophysical task-switching. *Animal Cognition, 17*, 619–631.

Baillargeon, R., & Hanko-Summers, S. (1990). Is the top object adequately supported by the bottom object? young infants' understanding of support relations. *Cognitive Development, 5*, 29–53.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated Measures Correlation. *Frontiers in Psychology, 8*.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package Version 0.999999-0. https://cran.r-project.org/package=lme4.

Berthier, N. E., DeBlois, S., Poirier, C. R., Novak, M. A., & Clifton, R. K. (2000). Where's the ball? Two- and three-year-olds reason about unseen events.. *Developmental Psychology, 36*, 394–401.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology, 56*, 73–102.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120*, 322–330.

Brosnan, S. F., & Hopper, L. M. (2014). Psychological limits on animal innovation. *Animal Behaviour, 92*, 325–332.

Burdett, E. R. R., McGuigan, N., Harrison, R., & Whiten, A. (2018). The interaction of social and perceivable causal factors in shaping 'over-imitation'. *Cognitive Development, 47*, 8–18.

Chen, Z., Siegler, R. S., & Daehler, M. W. (2000). Across the great divide: Bridging the gap between understanding of toddlers' and older children's thinking. *Monographs of the Society for Research in Child Development, 65*(2 Serial No. 261).

Costa, M. (2012). Territorial behavior in public settings. *Environment and Behavior, 44*, 713–721.

Davis, S. J., Schapiro, S. J., Lambeth, S. P., Wood, L. A., & Whiten, A. (2019). Behavioral conservatism is linked to complexity of behavior in chimpanzees (*Pan troglodytes*): Implications for cognition and cumulative culture. *Journal of Comparative Psychology, 133*, 20–35.

Deák, G., & Bauer, P. J. (1995). The effects of task comprehension on preschoolers′ and adults′ categorization choices. *Journal of Experimental Child Psychology, 60*, 393–427.

Eklund, A. (2016). beeswarm: The bee swarm plot, an alternative to stripchart. http://www.cbs.dtu.dk/~eklund/beeswarm/.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development, 10*, 483–527.

Gökçen, E., Petrides, K. V., Hudry, K., Frederickson, N., & Smillie, L. D. (2014). Sub-threshold autism traits: The role of trait emotional intelligence and cognitive flexibility. *British Journal of Psychology, 105*, 187–199.

Gönül, G., Takmaz, E. K., Hohenberger, A., & Corballis, M. (2018). The cognitive ontogeny of tool making in children: The role of inhibition and hierarchical structuring. *Journal of Experimental Child Psychology, 173*, 222–238.

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*, 87–92.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National academy of Sciences of the United States of America, 114*, 7892–7899.

Greve, W., & Thomsen, T. (2016). Evolutionary advantages of free play during childhood. *Evolutionary Psychology, 14*. 147470491667534.

Haun, D. B. M., Rekers, Y., & Tomasello, M. (2014). Children conform to the behavior of peers; other great apes stick with what they know. *Psychological Science, 25*, 2160–2167.

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*, e833.

Hopper, L. M., Schapiro, S. J., Lambeth, S. P., & Brosnan, S. F. (2011). Chimpanzees' socially maintained food preferences indicate both conservatism and conformity. *Animal Behaviour, 81*, 1195–1202.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children. *Animal Cognition, 8*, 164–181.

Hrubesch, C., Preuschoft, S., & van Schaik, C. (2009). Skill mastery inhibits adoption of observed alternative solutions among chimpanzees (*Pan troglodytes*). *Animal Cognition, 12*, 209–216.

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New Ideas in Psychology, 30*, 190–200.

Jacobson, S. L., & Hopper, L. M. (2019). Hardly habitual: Chimpanzees and gorillas show flexibility in their motor responses when presented with a causally-clear task. *PeerJ, 7*, e6195.

Keen, R. (2011). The development of problem solving in young children: A critical cognitive skill. *Annual Review of Psychology, 62* (1), 1–21.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition, 34*, 57–83.

Kirkham, N. Z., Cruess, L., & Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Developmental Science, 6*, 449–467.

Kirkham, N. Z., & Diamond, A. (2003). Sorting between theories of perseveration: Performance in conflict tasks requires memory, attention and inhibition. *Developmental Science, 6*, 474–476.

Legare, C. H., Sobel, D. M., & Callanan, M. (2017). Causal learning is collaborative: Examining explanation in social contexts. *Psychonomic Bulletin & Review, 24*, 1548–1554.

Loehr, A. M., Fyfe, E. R., & Rittle-Johnson, B. (2014). Wait for it Delaying instruction improves mathematics problem solving: A classroom study. *Journal of Problem Solving, 7*, 36–49.

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*, 284–299.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*, 19751–19756.

McGuigan, N., & Núñez, M. (2006). Executive functioning by 18-24-month-old children: Effects of inhibition, working memory demands and narrative in a novel detour-reaching task. *Infant and Child Development, 15*, 519–542.

Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition, 47*, 121–148.

Neldner, K., Mushin, I., & Nielsen, M. (2017). Young children's tool innovation across culture: Affordance visibility matters. *Cognition, 168*, 335–343.

Over, H., & Carpenter, M. (2013). The social side of imitation. *Child Development Perspectives, 7*(1), 6–11.

Pope, S. M., Fagot, J., Meguerditchian, A., Washburn, D. A., & Hopkins, W. D. (2019). Enhanced cognitive flexibility in the seminomadic Himba. *Journal of Cross-Cultural Psychology, 50*, 47–62.

Pope, S. M., Fagot, J., Meguerditchian, A., Watzek, J., Lew-Levy, S., Autrey, M. M., & Hopkins, W. D. (2020). Optional-switch cognitive flexibility in primates: Chimpanzees' (*Pan troglodytes*) intermediate susceptibility to cognitive set. *Journal of Comparative Psychology, 134*, 98–109.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: The task dynamics of the A-not-B error. *Psychological Review, 106*, 235–260.

Sneve, M. H., Grydeland, H., Rosa, M. G. P., Paus, T., Chaplin, T., Walhovd, K., & Fjell, A. M. (2018). High-expanding regions in primate cortical brain evolution support supramodal cognitive flexibility. *Cerebral Cortex, 29*, 3891–3901.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*, 605–632.

Spensley, F., & Taylor, J. (1999). The development of cognitive flexibility: Evidence from children's drawings. *Human Development, 42*, 300–324.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science, 348*, 91–94.

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition, 163*, 1–14.

Stengelin, R., Hepach, R., & Haun, D. B. M. (2020). Cross-cultural variation in how much, but not whether, children overimitate. *Journal of Experimental Child Psychology, 193* 104796.

Stoet, G., & Snyder, L. H. (2007). Task-switching in human and nonhuman primates: Understanding rule encoding and control from behavior to single neurons. In S. A. Bunge & J. D. Wallis (Eds.), *Neuroscience of rule-guided behavior* (pp. 227–254). Oxford, UK: Oxford University Press.

Watzek, J., Pope, S. M., & Brosnan, S. F. (2019). Capuchin and rhesus monkeys but not humans show cognitive flexibility in an optional-switch task. *Scientific Reports, 9*, 13195.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 2417–2428.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wilkinson, L. (1999). Dot plots. *The American Statistician, 53*, 276–281.

Winsler, A., Fernyhough, C., & Montero, C. (2009). *Private speech, executive functioning, and the development of verbal self-regulation*. Cambridge, UK: Cambridge University Press.

Wood, L. A., Kendal, R. L., & Flynn, E. G. (2013). Copy me or copy you? The effect of prior experience on social learning. *Cognition, 127*, 203–213.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology, 1*(2), 198–226.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11*, 37–63.

Zelazo, P. D., Muller, U., Frye, D., & Marcovitch, S. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development, 68*(3 Serial No. 274).