The Role of Causal Stability in Children's Active Exploration

Daniil Serko (serko@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development

Ny Vasil (ny.vasil@csueastbay.edu)
California State University, East Bay

Azzurra Ruggeri (ruggeri@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development

Abstract

Previous research documented adults' preference for stable causal relationships that do not vary in strength across backgrounds (Vasilyeva, Blanchard, & Lombrozo, 2018). In this study, we investigate the role of causal stability in guiding children's exploration behavior. We developed a computerized version of an active information-search paradigm to study how children dynamically explore different agents and backgrounds to learn more about their causal stability. Five- to seven-year-old children (n = 60) were presented with stable and unstable causes (i.e., causes with fixed or variable causal efficacy across backgrounds). We assessed children's causal attributions of outcomes and their exploratory behavior as they tried out previously observed and novel causes across previously observed and novel backgrounds. We find that children in this age range acknowledge causal instability in their causal attributions, and they become increasingly adept at tracking causal efficacy across multiple factors simultaneously (causes and backgrounds), but this does not translate into a blanket preference for exploring stable or unstable causes. We suggest a possibility that causal (in)stability guides exploration in more subtle and indirect ways and discuss the implications of our findings for the development of active exploration.

Keywords: active-learning; exploration; causal relationships; stability; backgrounds;

Introduction

Imagine you and your neighbor both decide to grow strawberries. Each of you plants the seeds in your own backyard and waits for the outcome. To your dismay, nothing happens in your garden — while your neighbor is harvesting bushels of juicy berries. You are puzzled: both you and the neighbor did the same thing, i.e., intervened on the same cause. Why the different outcomes? It turns out the chances that a cause will in fact produce the desired outcome—sweet and juicy strawberries—depend on several background variables, which might significantly increase or decrease the likelihood that the cause will produce an effect. For example, planting strawberry seeds is more likely to result in strawberries if the soil is acidic. If your backyard has alkaline soil, the same cause may not be as effective - so you end up with no berries.

This example illustrates the notion of causal *instability* across *backgrounds*. Philosophers have defined *stable* causal relationships as those that hold with similar strength across different backgrounds (where *backgrounds* can refer to any variable other than the cause and effect) (Woodward, 2006, 2010). Causal stability can be defined in terms of variability in causal *strength*. Various measures of causal strength exist, with important differences among them (Cartwright, 1989, 2009; Cheng, 2000; Cheng, Liljeholm, & Sandhofer, 2013; Cheng & Lu, 2017; Liljeholm & Cheng, 2007), but it can be glossed as a generalized measure of efficacy of a cause in generating an outcome, controlling for other factors. Suppose you place 8 seeds in your garden's soil and 8 in store-bought pre-fertilized soil. The number of strawberries you eventually obtain will reveal the causal strength in these specific backgrounds: more strawberries indicate higher causal strength. Assessing overall causal strength is important for selecting the most effective interventions from a range of possibilities (e.g., the best-producing variety of strawberries, or the best soil) (Meder, Mayrhofer, & Waldmann, 2014). But beyond that, assessing how much causal strength *varies* across backgrounds, its stability, can offer better guidance for predictions within and across backgrounds (Blanchard, Vasilyeva, & Lombrozo, 2018; Liljeholm & Cheng, 2007).

Empirical work with adults

Prior work has argued that it is particularly advantageous to keep track of stable causal relationships to generalize knowledge to new situations and contexts and demonstrated that adults prefer stable causal relationships over unstable ones in generalization and intervention (Blanchard et al., 2018; Lombrozo & Carey, 2006; Lombrozo, 2010; Woodward, 2006, 2010). For example, researchers presented participants with a hypothetical scenario of a supplement that is supposed to increase bone density and manipulated whether participants received information that the supplement had a stable effect on bone density (non-moderated group) or whether the effect varied depending on whether or not participants carried a specific gene (moderated group). Adults in the non-moderated group were more likely to agree with causal generalizations that the supplement increases bone density, even though the moderated and non-moderated causes had equal causal strength on average. Participants were also more likely to intervene on the stable cause (i.e., decide to take the pill to increase their own bone density, under the conditions of uncertainty about the background) (Vasilyeva et al., 2018). This research shows that adults prefer stable relationships over unstable ones when making causal generalizations across contexts and when deciding whether to intervene on a cause to produce an outcome.

However, most causal relationships are not stable, as they can be influenced—at least to some degree—by other variables. For example, even the relatively stable causal relationship of water boiling at 100 degrees Celsius is impacted

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

by the altitude at which the water is set to boil. Basing one's predictions on the assumption of causal stability alone might therefore be misleading (Cheng, 2000). Adults seem to monitor which relationships are unstable, and use this information to make rich inferences. For example, when participants learned that a pill's side effect (headaches) varied across treatment groups, they inferred that the causal relationship between pill and headaches might interact with another non-observable background factor, suggesting that people can infer the influence of additional background variables when they encounter *unstable* causal relationships (Liljeholm & Cheng, 2007).

Empirical work with children

For children, learning about and understanding causal relationships is particularly crucial, given that they are navigating the world with less data and experience than adults. Research indicates that young children are motivated causal learners: they spontaneously intervene on novel causal systems to infer the underlying causal structure, and form predictions about outcomes of their interventions (for reviews, see Gopnik and Wellman, 2012; Gopnik, 2012; Lapidow and Walker, 2019; see also Goddu and Gopnik, 2020).

Previous work suggests that children show some sensitivity to causal stability, and can use this information to guide their interventions (Cheng et al., 2013). In a recent study, children learned about farm and zoo animals that developed red dots and were treated with a specific diet: farm animals received a grain diet, and zoo animals received a grain-and-leaves diet. After children learned about the underlying probabilistic causal relationships, they had to choose which diets to administer to two new animals with red dots to make the dots disappear. The results show that children consider whether the effects observed in the animals can be attributed to grain alone or must involve an interaction of grain and leaves. In particular, when the grain diet had the same causal strength across the two contexts, children chose to feed the animals grain. In contrast, when the effects varied across contexts, and red dots disappeared more often with a grain-leaves-diet, they were sensitive to the differences in outcome due to the influence of the leaves and opted for the grain-and-leaves diet (Cheng, Sandhofer, & Liljeholm, 2022).

Moreover, children seem to be quite open to the possibility that causal relationships may not be stable over time. For example, Sumner et al. (2019) presented 4- to 12-year-old children and adults with a dynamic game environment in which they had to identify a reward-generating monster out of four options, across 80 rounds. In one condition, the reward-generating monster was switched after 40 trials, so participants had to explore the four options again to find it. Adults took much longer to detect the target monster's change than children. This highlights the learning advantages of prolonged exploration, as it allows to detect consequential environmental changes that moderate causal relationships (Sumner et al., 2019).

More generally, young children are sensitive to the impact of background factors on causal relationships. In particular, children as young as 2 years begin to understand that factors such as social norms and moral beliefs can impact the causal behavior they observe (Chernyak & Kushnir, 2014; Kalish & Shiverick, 2004; Rakoczy, Warneken, & Tomasello, 2008; Smetana, 1981; Turiel, 1983). By the age of 3, they begin to understand that emotions influence behavior (Harris, 1989; Lagattuta & Wellman, 2001), and at the age of 4 children explain variations in observed causal behavior citing situational factors as reasons (Seiver, Gopnik, & Goodman, 2013).

Taken together, this evidence suggests that even young children possess the cognitive competencies required to engage in reasoning about the stability of causal relationships. Yet, to our knowledge, no study to date has examined how young children explore stable and unstable causal relationships. Any real-world agents with limited resources must select what new data to pursue. Do children prioritize reducing uncertainty associated with unstable causes by selectively testing them in novel backgrounds? Or are children equally interested in collecting new data about stable and unstable causes—perhaps targeting a higher level uncertainty about whether the causes are indeed reliably stable or reliably unstable across a broad range of backgrounds? Answering these questions promise to expand our understanding of the factors shaping children's active learning about the world.

The present study

We investigated how 5- to 7-year old children respond to the stability of probabilistic causal relationships across contexts. This age range captures a critical period in the development of skills relevant for exploration behavior (i.e., attention, memory, executive functions) (Diamond, 2013; Roebers, Cimeli, Röthlisberger, & Neuenschwander, 2012). We developed a novel information-search task that allowed us to examine, on the one hand, how children dynamically explore different agents and backgrounds to learn more about their causal stability and, on the other hand, how the stability of the causal relationships under investigation impacts children's exploratory patterns. Children were introduced to two probabilistic causes (monsters) that were equated in average strength (probability of producing a lightning-bolt outcome) but varied in stability across backgrounds (planets). We measured whether children attributed outcomes to causes, backgrounds, or their combinations and whether they wanted to explore stable, unstable, or unknown causes in familiar and unfamiliar backgrounds.

The overall objective was to examine whether children are sensitive to causal stability and, if so, how it shapes their causal attributions and their exploration behavior. Specifically, first, how do children attribute outcomes to causes and backgrounds? If they notice and appreciate that backgrounds play an important moderating role in unstable relationships, they should attribute causal outcomes to the combinations of causes and backgrounds (monsters and planets) rather than to causes or backgrounds alone. Second, how do children select what causes to explore and intervene on, stable or unstable, in novel background contexts? Does their preference, if any, change with age? If children, like adults, prefer stability, they should intervene on stable causes (i.e., they should pick the stable monster). If they have not developed this preference yet, they should choose at chance. Yet another possibility is that children might have a preference opposite to that of adults, and favor unstable causes (i.e., they should pick the unstable monster). Third, what kind of information about causal relationships is a primary driver of children's exploration decisions, information about causal stability across backgrounds, or information about average causal strength (i.e. previously observed probability of an outcome associated with a given cause or background), or some other metric such as minimum or maximum causal efficacy observed so far?

Method

Participants

We recruited 60 children (27 female, $M = 74.70$ months; $SD = 11.57$ months; range: 60 to 95 months) through the participants' database of the Max Planck Institute in Berlin and tested them online via the Big Blue Button software. An additional 27 children were excluded from the analyses because they were too young (n=7), or failed to answer the comprehension check questions correctly (n = 20: 13 5-year-olds, 7 6-year-olds). Parents signed an informed consent form, and children agreed by giving verbal assent. The study was preregistered via OSF (Link: `https://osf.io/2xb98`) and approved by the ethics committee of the Max Planck Institute for Human Development ethics committee in Berlin (N-2021-01). The sample size was determined by conducting a simulation-based a-priori power calculation, to detect a hypothesized effect of .15 (Cohen's *d*) for an interaction of age and task with .80 power and an alpha significance level of .05. The initially registered age range was 5 to 6 years old, but it was expanded to include 7-year-olds prior to data collection.

Design, Materials and Procedure

Participants sat next to their parents in front of a computer and were introduced to a game via screen share. The study consisted of three phases: familiarization, exploration without feedback, and free-exploration with feedback. An attribution question was presented twice, first after the familiarization phase, and second after the exploration phase.

Familiarization Children were introduced to a space-themed game in which they observed two types of causes (turquoise/yellow monsters) generating a probabilistic outcome (energy in the form of lightning bolts) on different backgrounds (red/blue planets). For example, on one learning trial, a group of 8 yellow monsters traveled to a red planet. Upon landing, some of them produced energy (visualized as overlaid lightning bolts), and children were asked to count the lightning bolts (with encouragement to re-count if they made an error). *Stable* monsters produced lightning bolts with a rate of 5/8 on both planets. *Unstable* monsters produced energy with a rate of 3/8 on one planet and 7/8 on the other planet. Importantly, on average, both the stable and unstable monsters produced the same amount of energy (10 out of 16 observations, see Planets 1 and 2 in Table 1). By the end of the familiarization phase, a child would have seen fewer lightning bolts on one of the planets (8 bolts total, composed of 5 and 3 bolts produced by the stable and unstable monsters, respectively) compared to the other planet (12 bolts total, generated by the stable (5) and unstable (7) monsters); we refer to these planets as *low-energy* and *high-energy* planets, respectively (monster and planet colors counterbalanced across participants). Once children had observed one group of monsters visiting both planets, they completed a comprehension check, indicating whether the monsters produced the same energy on both planets or more energy on one of the two planets. Children completed one comprehension check for the stable and the unstable monsters, respectively. Children who failed these comprehension checks repeated the familiarization phase. If they failed to answer the comprehension checks after three familiarization rounds, they were excluded from the sample ($n = 20$; 10 female; $M = 70.05$ months; $SD = 6.35$). At the end of the familiarization phase, children were presented with a summary slide showing both monsters next to both planets, with the number of energy bolts they had produced on each planet.

Attribution questions After the familiarization phase, children completed the first causal-attribution task. They were presented with three statements attributing the outcome ("lightning bolts happen...") either to the causes ("because of the monsters"), or to the backgrounds ("because of the planets"), or to both (" because of the monsters and the planets"); each claim appeared in a speech bubble of a uniquely-colored unicorn. Children selected the unicorn they thought was right. At the end of the study, children were again presented with the same attribution question (unicorn colors counterbalanced within and between participants).

Exploration phase without feedback Children made six decisions, each involving a choice between two options. In three decisions, they chose between two causes, and in the remaining three they chose between two backgrounds (order pseudo-randomized). They did not receive any feedback about the outcomes of their choices. Out of the six decisions, two involved a novel element: either a novel cause or a novel background. The key decision trial (*novel planet*) assessed children's preference for intervening on a stable or unstable cause under conditions of background uncertainty: a child was presented with a novel planet (unfamiliar background), and was asked to decide whether to send a stable or unstable monster to this planet. A preference for exploring (un-)stable causes would manifest in selecting the respective monster in this task. (Note that, like many real-world decisions, this task can be construed as having elements of exploitation—applying prior knowledge to generate a desired outcome—and exploration—learning how a cause functions in a previously unexplored background. However, two features of this task maximize its exploratory character: first, the two causes were equated in average causal strength, such that the expected probability of generating an outcome (lightning bolts) could not offer guidance for selecting one monster over the other in a new background; second, children did not receive any rewards or prompts to produce a high number of energy bolts at any point throughout the task; we say more on this in the Discussion.)

The second key decision trial examined whether children tracked the average causal strength of familiar backgrounds and used this information to inform their interventions. On this trial (*novel monster*), children were presented with a new monster and could decide whether they wanted to send it to the low or high-energy planet. If children aim to maximize the chances of producing the outcome based on the average causal strength they should pick the high-energy planet. Again, no incentives or prompts to produce a high number of energy bolts were given.

The remaining four decisions involved familiar combinations of causes and backgrounds that children had previously encountered during the familiarization phase. These questions allowed us to assess the extent to which children's choices were driven by general preferences for (in-)stability vs. by maximizing expected outcomes based on the previously observed average causal strength of each variable. On two *old planet* trials, children saw one planet (either the low or the high-energy planet) and chose whether to send there the stable or unstable monsters. If a preference for stability drives children, they should pick the stable monsters on both trials; if they prefer instability, they should consistently send the unstable monsters. If the causal strength instead drives their preferences, they should send the stable monster to the low-energy planet and the unstable monster to the high-energy planet to maximize outcomes. On the remaining two *old monster* trials, children saw one monster type (either the stable or unstable) and chose whether to send them to the low or the high-energy planet. Because causal stability offers no grounds for preferring one planet over the other, children relying on stability alone should choose at chance. If they rely on causal strength instead, they should always pick the high-energy planet. Since children received no feedback about the outcomes of their choices (i.e., they did not get to see what happened after the selected monsters traveled to the planets, etc.), they did not accumulate in this phase any new data about causal stability or average causal strength.

Free-exploration phase with feedback Children tried out different combinations of familiar and novel causes across familiar and novel backgrounds and observed the outcomes. We wanted to mirror an everyday situation where children encounter various causal relationships across different backgrounds and have the chance to explore the relationships freely, without explicit guidance or incentives. Therefore, we did not incentivize or encourage them to generate as much energy as possible. With this approach children may be more likely to explore and engage with the task in a more open-ended way, which can reveal their underlying cognitive processes and strategies without being narrowly focused on a particular goal. Children could choose between five different types of monsters (see Figure 1), send each monster type (in groups of eight) to one of six different planets of their choice, and observe how many of the eight monsters generated energy bolts on a particular planet. This task allowed us to examine what types of causes children were most interested to explore (and re-explore).

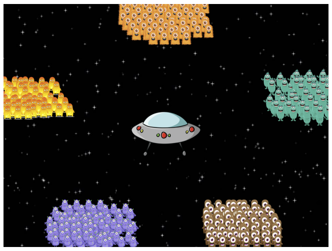

Figure 1: Screenshot from the free-exploration phase with feedback, with five cause options. Once a child selected one monster type, a group of eight monsters boarded the space shuttle and the child proceeded to select one planet to send the monsters to.

The five monster options included the stable and unstable monsters from the familiarization phase and three new monster types (monsters generated energy following one of the patterns shown in Table 1). The familiar monsters continued to produce energy displaying the previously observed patterns: the *old stable* monsters always generated energy with the rate of 5/8; the *old unstable* monsters alternated between generating 3/8 and 7/8 energy bolts across planets. The three new monster types included the *new stable low* monsters, which produced 1 energy bolt across the 8 monsters on all planets; the *new stable high* monsters produced 7 energy on all planets; finally, the *new unstable* monsters alternated between 0, 2 and 8 energy bolts depending on the planet they visited. Energy-production patterns were counterbalanced across monsters. The planets included the familiar red and blue planets and four novel planets (including one novel planet featured in the Exploration without feedback task). At the end of each exploration round, children counted the energy bolts produced. If they miscounted, they were encouraged to recount. Every four rounds, we asked children whether they wanted to "continue or stop" playing the game (wording counterbalanced within subjects: "stop or continue").

Table 1: Free-exploration phase with feedback: number of monsters generating energy bolts (out of 8 monsters), across different planets.

	Planet					
Monster	P1	P2	P3	P4	P5	P6
Stable	5	5	5	5	5	5
Unstable	3	7	3	7	3	7
New low	1	1	1	1	1	1
New unstable	8	2	0	8	2	0
New high	7	7	7	7	7	7

Results

Attribution Questions

On the first attribution question, the majority of the children (48.33%) attributed the outcome (energy bolts) to the combination of causes and backgrounds, i.e., monsters and planets, rather than to causes, i.e., monsters (26.67%), or backgrounds, i.e., planets (25.00%, $\chi^2(2) = 6.100$, $p = .047$). On the second attribution question, children's choices followed the same ordering, with 38.98% of children attributing the outcome to the combination of causes and backgrounds, 37.29% attributing it to causes, and 23.73% attributing it to backgrounds; however, these differences were not significant, $\chi^2(2) = 2.475$, $p = .290$. The difference between the two rounds of attribution questions was not significant ($\chi^2(2) = 1.399$, $p = .497$). To investigate whether the propensity to attribute outcomes to interactions changes with age, we re-coded responses into a binary variable, attributions to the cause x background interaction vs. a single factor (either causes or backgrounds). Age in months did not significantly predict these responses in a logistic regression, $p = .707$, $OR = 0.991$ [0.946 – 1.039].

Exploration phase without feedback

We began by examining choices on the *novel planet* trials, where children had the option of intervening on either the stable or unstable cause in a novel background (i.e., sending either stable or unstable monsters to a new planet they had no prior information about). Overall, 45% of the children sent the *stable* monsters to the new planet, which did not significantly differ from chance (50%, $p = .519$, exact binomial test). A logistic regression predicting choices from age revealed no developmental change ($p = .238$, $OR = 1.028$ [0.982 – 1.076]) (see Figure 2).

We then turned to the two *old planet* trials, where children selected which monsters to send to the low and high-energy planets. This allows us to assess how children apply the prior evidence they gathered during the familiarization phase about causes and backgrounds in designing interventions. On average, when presented with the low-energy planet, half of the children chose the stable and the other half the unstable monsters (50%, $p = 1.000$, exact binomial test). When presented with the high-energy planet, 57% of the children preferred to send the stable monsters, which did not differ from chance ($p = .366$, exact binomial test). A logistic regression predicting monster choice from planet type (low vs. high-energy planet) and children's age in months revealed that age alone ($p = .305$, $OR = 1.024$ [0.979 – 1.071]) did not predict children's decisions. However, the type of planet presented ($p = .067$, $OR = 97.194$ [0.729 – 12958.617]) marginally predicted children's decisions. Most importantly, the interaction of age and the presented planet was significant ($p = .050$, $OR = 0.937$ [0.878 – 1.000]); as shown in Figure 2, with age children became more selective, sending the unstable monsters more to the low-energy planet, and the stable monsters more to the high-energy planet.

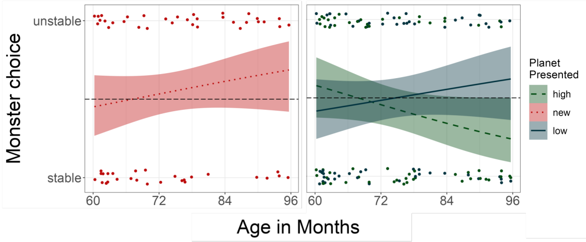

Figure 2: Children's choices between stable vs. unstable causes (monsters) in the exploration without feedback phase, when presented with the novel background (planet) *(left panel)*, or when presented with the old backgrounds (high-energy vs. low-energy planet) *(right panel)*. Each dot represents a child's choice between the stable and unstable causes. The lines indicate a fitted logistic regression.

Next, we examined responses from the *novel monsters* trial, where children chose to send new monsters they had no data about to either the low or high-energy planet. We found no evidence that children relied on average causal strength in this decision: 57% of children sent the new monsters to the high-energy planet, exact binomial test against chance 50%, $p = .366$. A logistic regression revealed that their choices did not vary with age ($p = .266$, $OR = 0.974$ [0.931 – 1.020]).

Children's choices on the two *old monsters* trials, where they were presented with either the stable or the unstable monsters and selected a planet (low or high-energy) to send each monster group to, also did not reveal significant preferences. When presented with the stable monsters, 55% of the children sent them to the high-energy planet ($p = .519$, exact binomial test). When presented with the unstable monsters, 55% decided to send them to the high-energy planet ($p = .519$). A logistic regression predicting planet choice from monster type (stable vs. unstable) and children's age in months revealed no effects of age ($p = .616$, $OR = 0.989$ [0.946 – 1.034]) or the type of monster ($p = .846$, $OR = 1.606$ [0.014 – 189.426]), and the interaction of age and monster type was not significant ($p = .844$, $OR = 0.994$ [0.933 – 1.059]).

Free-exploration phase with feedback

On average, children performed 8.26 (SD = 6.89) rounds of explorations. Most children (78%) tried at least some monsters more than once, and 38% of children re-explored all monsters. To investigate this further, we specified monster type as a predictor of whether children re-explored it. Overall, monster type predicted re-exploration behavior: the new low monster (always producing 1 energy bolt) ($p = .027$, $OR = 3.725$, [1.163 – 11.936]) and the new unstable monster (producing 0/2/8 energy bolts alternating) were significantly more likely ($p = .006$, $OR = 5.287$, [1.600 – 17.475]) to be re-explored than the old unstable monsters.

Discussion

We investigated whether 5- to 7-year-old children's active-exploration strategies are sensitive to the (in-)stability of causal relationships. We find that overall, in this age range, children can already appreciate the interactive nature of causal relationships. After they were presented with evidence that some causes act differently in different backgrounds, they attributed outcomes to a combination of causes and backgrounds rather than to either causes or backgrounds in isolation.

We find evidence that, with age, children use prior evidence more in designing interventions involving familiar combinations of causes and backgrounds that lead to low outcomes. In the exploration without feedback phase, older children tended to be more selective, using the *unstable* cause in the background context where this cause had been previously less effective (3/8, which is lower than the stable cause's performance of 5/8) and switching to intervening on the *stable* cause in the background where this cause had been less effective during the familiarization phase (5/8, which is lower than the unstable cause's performance of 7/8). This reveals an increasing capacity to track and integrate information about causes and backgrounds to guide exploration decisions.

While the results reported above are promising, we failed to find evidence that children's active exploration is directly guided by causal stability or by a preference for causal strength. This is surprising, given the prior empirical evidence that preschoolers are sensitive to stable causal relationships (Cheng et al., 2022). Instead we found that children's active exploration targeted causes with the lowest minimum observed causal efficacy. For example, in the exploration with feedback phase children repeatedly explored monster groups generating zero or one energy bolts in at least some backgrounds. This could be due to a variety of factors. One possibility is that children were simply drawn to low-energy outcomes for reasons beyond our study setup—perhaps they are budding environmentalists, who had learned that saving energy is crucial from their parents or at school. Another possibility is that the interest in exploring ineffective causes stems from children's expectations that these causes are unstable across backgrounds. Perhaps they were trying to find a background where these causes would turn out highly effective (looking for a "jackpot"). One way to examine this further would be to compare the exploration behavior of children who had and had not been exposed to unstable relationships beforehand. This lies beyond the scope of this paper. In our study, all children had witnessed unstable relationships in the familiarization phase, which likely made the possibility of contextual variability in causal strength more salient to all of them, which could make them seek fortuitous backgrounds for ineffective causes.

Children's general lack of preference for stable or unstable causes in exploration tasks can reflect several things. First, children may have been uncertain about what would be most beneficial to learn in this task. We did not offer incentives for generating outcomes, and the valence of the outcome was left ambiguous (we did not offer any guidance on whether it is better to produce as much energy as possible or to save energy); this openness could have resulted in high variability across children in our sample in terms of what each of them was trying to discover or achieve during exploration. Second, this may have been a challenging task with too many choice options. Since one must explore a cause in at least two backgrounds to know whether it is stable or unstable, the task of determining stability for five causes may have exceeded children's capacity. Reducing the free-exploration phase to four monsters and planets and ensuring all children gathered enough data about all causes might provide more precise insights in a future study. Third, our findings might mean that causal (in)stability does not matter in the context of exploration tasks (although Sumner, 2019, suggests otherwise). At this point, we do not have adult data for comparison; it is possible that while adults show stability preference in some tasks, they do not rely on it in exploration.

We are preparing a set of follow-up studies that address these possibilities by i) incentivizing children for each energy bolt they produce; ii) clearly stating that it is favorable to produce energy, for instance, by asking children to help the monsters restart their space shuttle by collecting as many energy bolts as possible; iii) running the study with older children (8- to 10-years-old) and adults and comparing the results between and within these age samples and the current sample; iv) requiring each child to explore all monsters at least twice, providing access to stability information for exploration decisions; v) implementing a computational approach to compare children's behavior against computational agents with a perfect preference for stability, instability, and causal strength.

In sum, children show signs of sensitivity to causal instability between 5 and 7 years of age (as revealed by their causal attributions). They become increasingly adept at tracking causal efficacy across multiple factors at the same time (causes and backgrounds), but they do not yet put this understanding to use to guide their exploration behaviors; at least, they do not show a blanket preference to explore stable or unstable causes; the possibility that causal (in)stability guides them in more subtle and indirect ways remains open and will be assessed in future studies.

References

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018). Stability, Breadth and Guidance. *Philosophical Studies*, *175*(9), 2263–2283. doi: 10/grm7qf

Cartwright, N. (1989). *Nature's Capacities and their Measurement*.

Cartwright, N. (2009, October). What are randomised controlled trials good for? *Philosophical Studies*, *147*(1), 59. doi: 10/d9zf9c

Cheng, P. (2000). Causality in the mind: Estimating contextual and conjunctive power. In *Explanation and cognition* (pp. 227–253). Cambridge, MA, US: The MIT Press.

Cheng, P., Liljeholm, M., & Sandhofer, C. (2013). Logical consistency and objectivity in causal learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35).

Cheng, P., & Lu, H. (2017). Causal invariance as an essential constraint for creating representation of the world: Generalizing the invariance of causal power. *The Oxford handbook of causal reasoning*, 65–84.

Cheng, P., Sandhofer, C. M., & Liljeholm, M. (2022). Analytic Causal Knowledge for Constructing Useable Empirical Causal Knowledge: Two Experiments on Pre-schoolers. *Cognitive science*, *46*(5), e13137. doi: 10/grp9fh

Chernyak, N., & Kushnir, T. (2014). The self as a moral agent: Preschoolers behave morally but believe in the freedom to do otherwise. *Journal of Cognition and Development*, *15*(3), 453–464. doi: 10/grp9fg

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. doi: 10/b2m2

Goddu, M. K., & Gopnik, A. (2020). Learning what to change: Young children use "difference-making" to identify causally relevant variables. *Developmental Psychology*, *56*(2), 275. doi: 10/grpbg7

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627. doi: 10/gc3kcn

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085. doi: 10/f4chc4

Harris, P. L. (1989). *Children and emotion: The development of psychological understanding.* Basil Blackwell.

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, *19*(3), 401–416. doi: 10/dqch48

Lagattuta, K. H., & Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Development*, *72*(1), 82–102. doi: 10/bf3fpm

Lapidow, E., & Walker, C. M. (2019). Does the intuitive scientist conduct informative experiments?: Children's early ability to select and learn from their own interventions. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 44, pp. 2085–2091).

Liljeholm, M., & Cheng, P. (2007). When Is a Cause the "Same"?: Coherent Generalization Across Contexts. *Psychological Science*, *18*(11), 1014–1021. doi: 10/fb26x7

Lombrozo, T. (2010, December). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. doi: 10/fvx8g3

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. doi: 10/fsdg8x

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277. doi: 10/f6ds5k

Rakoczy, H., Warneken, F., & Tomasello, M. (2008, May). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881. doi: 10/br6p85

Roebers, C. M., Cimeli, P., Röthlisberger, M., & Neuenschwander, R. (2012, December). Executive functioning, metacognition, and self-perceived competence in elementary school children: An explorative study on their interrelations and their role for school achievement. *Metacognition and Learning*, *7*(3), 151–173. doi: 10/gfvw2c

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child development*, *84*(2), 443–454. doi: 10/f4s8pt

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child development*, 1333–1336. doi: 10/d7cskx

Sumner, E., Li, A. X., Perfors, A., Hayes, B., Navarro, D., & Sarnecka, B. W. (2019). The Exploration Advantage: Children's instinct to explore allows them to find information that adults miss.
doi: 10/grqb4n

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296. doi: 10/gdp5bc

Woodward, J. (2006). Sensitive and Insensitive Causation. *The Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318. doi: 10/b2wbmr