Calculating probabilities from imagined possibilities: Limitations in 4-year-olds

Brian Leahy,[†,1] **Vicente Vivanco,**[†,1] **Sam Cheyette,**[1] **Kevin A Smith,**[1]
Lucy White,[2,3] **Roman Feiman**,[2] **Laura Schulz**,[1] **Joshua B Tenenbaum**[1]
[1]Massachusets Institute of Technology, [2]Brown University, [3]University of Bath
[†] brianlea@mit.edu, vvc@mit.edu

Abstract

Adults can calculate probabilities by running simulations and calculating proportions of each outcome. How does this ability develop? We developed a method that lets us bring computational modeling to bear on this question. A study of 40 adults and 31 4-year-olds indicates that unlike adults, many 4-year-olds use a single simulation to estimate probability distributions over simulated possibilities. We also implemented the 3-cups task, an established test of children's sensitivity to possibilities, in a novel format. We replicate existing 3-cups results. Moreover, children who our model categorized as running a single simulation on our novel task show a signature of running a single simulation in the 3-cups task. This signature is not observed in children who were categorized as running multiple simulations. This validates our model and adds to the evidence that about half of 4-year-olds don't evaluate multiple candidates for reality in parallel.

Keywords: Possibility; probability; modal concepts; computational modeling

Introduction

Adults can compute probabilities in many ways. One way involves estimating proportions in observable populations: if we see an urn that holds 80% red balls, we know that a randomly-drawn ball will most likely be red. This capacity is evolutionarily ancient (Gallistel, 1990) and develops early (Saffran, Aslin, & Newport, 1996; Xu & Garcia, 2008). Second, adults can use their intuitive physics or other mental models to run several simulations; the proportion of simulations with each outcome can be treated as the probability of that outcome (Battaglia, Hamrick, & Tenenbaum, 2013). The current paper addresses the ontogenesis of this capacity. We evaluate concrete hypotheses about how children and adults use simulation to generate probability distributions.

Preschoolers often make unwise decisions when faced with multiple possibilities. In the 3-cups task (Mody & Carey, 2016) children see 3 cups: a pair and a singleton. A prize is hidden in the singleton, and another prize is hidden in the pair–the child can't tell which cup. The child chooses one cup and keeps its contents. Choosing the singleton guarantees a prize. Older 2-year-olds (30-36 months) choose the singleton half the time; analyses of how errors distribute over participants show that all older 2-year-olds choose the singleton with probability .5 (Leahy, Huemer, Steele, Alderete, & Carey, 2022). Random choice should yield 1/3 choice of the singleton, which is not observed; why would individual children choose the singleton with probability .5? Leahy & Carey (2020) propose that these children lack *modal concepts*, and hence cannot model multiple versions of a single reality. Instead they deploy *minimal representation of possibility*. They run just one simulation, in which each prize lands in a cup. But since they only simulate once, they do not understand that the prize in the pair *might* be in either cup. They take its location to be determined. So they choose randomly between the two locations that they believe hold a prize (the singleton and one member of the pair). Over many trials, this yields 50% choice of the singleton and 25% choice of each member of the pair, as observed in older 2-year-olds. With age, a group of children who more reliably choose the singleton emerges and grows. At age 4, only about half of children seem to understand that there are multiple possibilities that need to be taken into account (Leahy, 2023). The remainder choose the singleton cup with probability .5. A similar proportion of children succeed on the Y-shaped tube task (Redshaw & Suddendorf, 2016). A ball is dropped into a tube shaped like an upside-down "Y"; 36-month-olds who want to catch the ball rarely cover both possible openings (Redshaw & Suddendorf, 2016), especially if there is a third, impossible opening that they must avoid. About half of 48-month-olds identify the two possible openings (Leahy, 2024).

There are many objections to the hypothesis that only about half of 4-year-olds (and fewer younger children) have modal concepts (Cesana-Arlotti, Kovács, & Téglás, 2020; Cesana-Arlotti & Halberda, 2024; Alderete & Xu, 2023; Turan-Küçük & Kibbe, 2024, 2025; Andreuccioli et al., 2024; Brody, Mazalik, & Feiman, 2024). Work on both sides of this debate typically uses binary response measures (e.g., does the child choose the singleton cup or not?), which make it easier to extract interpretable data from preschoolers. However, binary measures do not support computational modeling that distinguishes between fine-grained hypotheses. The current paper introduces a richer dependent measure that allows us to more precisely test quantitative, mechanistic accounts of how children use simulation in their intuitive physics engine to evaluate possibilities. Following Gerstenberg, Siegel, and Tenenbaum (2018), we created virtual "Plinko" boxes (Fig 1) where small balls fell through a series of obstacles before landing in a bin at the bottom of the screen. Children could put virtual "cushions" into the bins. When a ball strikes a cushion, it plays a jingle, grows, and turns golden; the cushion disappears and the ball keeps falling. So the ball can hit multiple cushions and gather multiple rewards.

We developed a novel probability task and a version of the

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

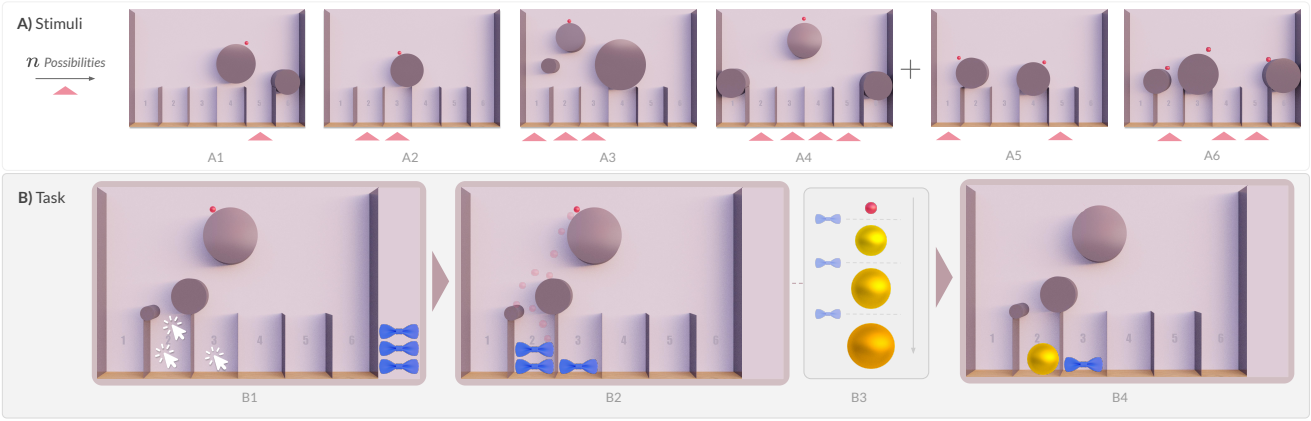

Figure 1: Probability task. A. Example stimuli. Red triangles indicate bins where the ball might land. In panels A1-A4 there is 1 ball. It either follows a determinate path (A1), so there was only 1 place for it to land, or an indeterminate path (A2-A4), with 2-4 bins for it to land in. In "2 ball, deterministic" trials (A5) there were 2 places for balls to land. In "3 ball, deterministic" trials (A6) there were 3 places for balls to land. Deterministic trials are controls: When there is only one possible bin (A1), do participants stack all their cushions there? When there are two places where balls where balls will definitely land (A5), do participants spread their cushions out over both possible bins? B. The task, illustrated with a "3 possibilities" trial (cf. A3). (B1) Participants were asked 3 times "Where would you like to put this cushion?". (B2) After each question, the experimenter placed a cushion in the indicated bin, yielding a distribution of cushions. (B3) The experimenter dropped the ball, which grew and changed color with each cushion it struck. (B4) Result: If the ball has landed on 1 or more cushions, it is bigger and golden.

3-cups task. In the probability task (Fig 1), children saw 15 trials where they placed 3 cushions in bins. When there was only one bin for the ball to land in (Fig 1A1; call these "deterministic" trials) stacking all cushions in that bin guarantees the greatest reward. In what we will call "probabilistic" trials, there were 2, 3, or 4 possible bins for the ball to land in (Fig 1A2-A4). With multiple possible bins, stacking cushions in one bin risks hitting no cushions at all. But children who deploy minimal representations of possibility will be insensitive to the difference between probabilistic and deterministic trials, since they only see one place for the ball to go. Spreading cushions into possible bins will be evidence that the child runs more than one simulation. Stacking cushions up will–with caveats, below–be evidence that the child deploys minimal representation of possibility.

In our modified 3-cups task (Fig 2), children placed one cushion. There were two balls. One ball followed a determinate path while the other ball could fall into either of two bins. Reliably placing the cushion under the ball that follows a determinate path indicates differentiating the sure thing from the mere possibilities. But children who deploy minimal representations of possibility should place the cushion under the ball that follows a determinate path with probability .5.

We test 3 hypotheses about how many simulations each child runs in the probability task. For each participant, we calculate the probability that they (1) run a single simulation and use that single result to guide their cushion placement (i.e., minimal representation of possibility); (2) run one simulation per cushion, using that simulation to place that cushion, repeating for each cushion; or (3) run enough simulations to have a good sense of the probability of the ball landing in each bin. Children who only run one simulation will tend to stack their cushions up; the bin they choose to stack in will be proportional to the bin's probability. So if there are two possible bins and one bin is twice as likely as the other, over many trials they will stack in the high probability bin twice as often as in the low probability bin. Children who simulate once per cushion will place cushions in bins in proportion to the bin's probability. For example, if there are two possible bins and one bin is twice as likely as the other, they most likely put two cushions in the higher-probability bin and one cushion in the lower-probability bin. The behavior of children who run many simulations will depend on how risk-seeking or risk-averse they are. Risk-seeking children will tend to stack their cushions up. However, they will stack their cushions in the highest probability bin, revealing their sense of the probabilities. This distinguishes gamblers who know the probabilities from children who run a single simulation. Risk-averse children will tend to spread their cushions out, prioritizing covering all the possibilities over getting more rewards.

The 3-cups task tests for convergence across tasks. Children who run a single simulation on the probability task are expected to choose wisely on the 3-cups task with probability .5. Children who run more simulations should perform better on the 3-cups task than those who run a single simulation.

Methods

Participants

Participants were 40 adults, recruited on Prolific, and 31 4-year-olds (range = 4.15-4.98, mean = 4.58, sd = 0.25), re-

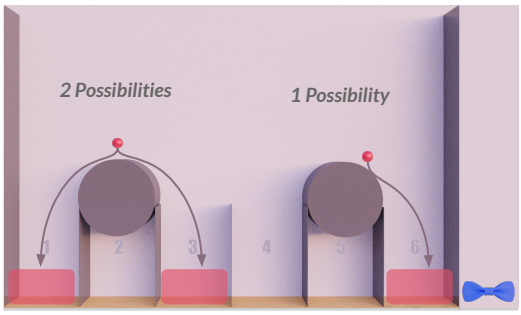

Figure 2: 3-cups trials: A conceptual replication of Mody & Carey's (2016) 3-cups task. Participants have one cushion to place. There are two balls. One ball follows a determinate path; the other is equally likely to fall into either of two bins.

cruited through childrenhelpingscience.com. Adults participated online in an unmoderated session. Children were tested in a moderated zoom session. Adults only completed the probability task, because in piloting adults had no variance on the 3-cups task (100% performance). Fourteen children were tested and excluded: 6 for withdrawing consent, 4 for equipment failure, and 4 for failing the training criteria.

Design

The *3-cups task* and the *Probability task* were blocked and counterbalanced. The test-phase of the probability task was presented in one of two pseudorandom orders, counterbalanced. All probability trial types were interleaved but biased so that trials with more possible bins tended to come later in the study. This gave children time to warm up to the game's mechanics before getting to the most difficult problems.

Stimuli and Procedure

Probability Task After learning the basic mechanics, children were asked to "make sure every ball gets big and golden". Then a training phase began, using only deterministic trial types (Fig 1A1, A6). Children placed 3 cushions into bins. First came a 1-ball deterministic trial (Fig 1A1). The experimenter indicated the top cushion in the sidebar and asked, "Where should I put this cushion?". This was repeated for each cushion. If the child did not place all 3 cushions in the correct bin, the experimenter used scripted prompts to help the child to understand why they should put all the cushions in the bin where the ball would land. Next came a 3-ball deterministic trial (Fig 1A6), following the same procedure. If the child did not place a cushion under each ball, the experimenter used scripted prompts to help the child understand why they should put a cushion under each ball. Finally, children who needed prompting were given a second trial of each trial type they needed prompting on. If they still needed prompting they were excluded.

In the test phase children saw 3 trials each of 5 trial types (Fig 1A1-A5). As in the practice phase, the experimenter indicated the top cushion in the sidebar and asked, "Where should I put this cushion?", placed the cushion, and repeated for all 3 cushions. Once all cushions were placed the ball was dropped.

3-cups task The 3-cups task conceptually replicated tasks that check whether children differentiate sure things from mere possibilities (Mody & Carey, 2016; Leahy, 2023). Children placed one cushion. In the training phase, they first saw a trial where a single ball would follow a determinate path. They were asked what bin the ball would fall in; once they identified the correct bin, the cushion was placed there and the ball was dropped. Next was a trial where a single ball had an indeterminate path. The experimenter said, "This ball might land here or here (indicating each possible bin). Let's watch and see which bin the ball lands in!" Once the ball had landed, participants were asked which bin the ball was in. The next trial was identical except the ball went into the other bin, showing participants that the ball could go both ways. If children made a mistake indicating which bin the ball was in, the process was repeated. Children who never made two correct responses in a row were excluded. No cushions were placed on these trials, as we did not want to encourage guessing. In the test phase there were two balls. One had a deterministic path; the other could land in two places (Fig 2). The experimenter asked, "Where do you want to put this cushion?", placed the cushion, and dropped the ball. There were 5 trials.

Results

Probability Task

We coded responses into three categories. "Stacking" means placing all three cushions in one possible bin. "Spreading" means placing cushions in more than one possible bin, and none in an impossible bin. "Other" means any other response, and is equivalent to placing at least one cushion in an impossible bin. Figure 3 shows the descriptive results. In One ball, deterministic trials, children stacked their cushions on 76% of trials. In two ball, deterministic trials, children spread out their cushions on 66% of trials. These results show that children see the value of stacking when there is one place for a ball to land, and of spreading when there are two places for a ball to land. On probabilistic trials, children stacked their cushions up 52% of the time and spread 32% of the time.

Aggregate comparison At the group level, children and adults distributed their cushions similarly. A model where adult and child aggregate cushion placements were drawn from the same distribution fit better than one where they come from separate distributions, for each of the probabilistic stimuli (1-ball, 2-, 3-, and 4-possibility trial types: mean ΔBIC = 6.9, 14.0, and 18.1 respectively).

Model The child and adult distributions do not differ in aggregate, but results could distribute over participants in different ways. We formulated a Bayesian hierarchical model to jointly infer both the strategies that participants used to generate their responses and the frequency of those strategies in the

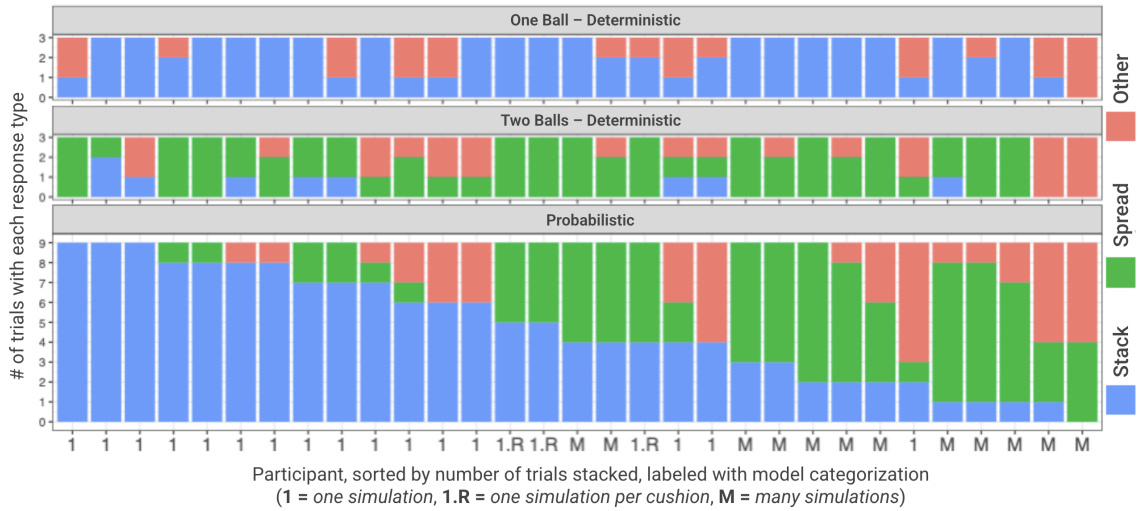

Figure 3: Test-phase descriptive results (children). Top row: results from the 3 1-ball deterministic trials. Middle row: results from the 3 2-ball deterministic trials. Bottom row: results from the 9 probabilistic trials. Participants were sorted by the number of times they stacked cushions in the Probabilistic trials; bars were labeled with the model categorizations of each participant.

population. We tested 3 hypotheses about how participants draw on the intuitive physics engine to calculate the probability of the ball landing in each bin: (1) run a single simulation and take the result to be the fact of the matter; (2) run a single simulation for each cushion; (3) run enough simulations to have a good sense of how probability is distributed.

We assumed that the aggregated adult distribution of cushions in each trial matched the ground truth of how probability was actually distributed over the bins in that trial. We then modeled running one simulation in the intuitive physics engine as drawing a single sample from the ground truth (adult) distribution. We will call the collection of samples drawn from the ground truth distribution the participant's "mental distribution". The mental distribution was then passed through a softmax function with temperature τ, allowing us to model each participant's risk-aversion or risk-seekingness. Finally, we assumed that participants place cushions by sampling 3 times from the softmaxed mental distribution, placing a cushion in each sampled location, with a noise parameter α determining the probability of placing a cushion randomly.

To help intuitively understand this model, we illustrate how it teases apart our 3 hypotheses and show how this depends on the model parameters. A child who runs many simulations will have a mental distribution that matches the ground-truth probability distribution. But this does not tell the child where to put their cushions. Moreover, some people are risk-seeking (they try to get more cushions) while others are risk-averse (they make sure they get at least one cushion). We model this with the parameter τ of the softmax function. In most cases (see the next paragraph for an exception), applying a softmax with low τ emphasizes the differences between the various possibilities, increasing the probability of high-value possibilities and decreasing the value of low-value possibilities. A high τ smooths out the differences between possibilities, increasing the probability of low-value possibilities and decreasing the probability of high-value possibilities. We assume that children sample from the soft-maxed mental distribution in deciding where to put their cushions. Children who are estimated to have a low τ-value will tend to get 3 samples from the highest-probability bin, and so will place all 3 cushions in the highest probability bin, modulo the probability α of placing a cushion in a random bin. Children with a high τ-value will tend to get samples from multiple bins, and so will spread their cushions out; there will also be some probability α that they place their cushion in a random bin.

Children who run a single simulation have a mental distribution where all simulations (all 1 of them) result in the ball landing in the same bin: a distribution with no variance. Applying a softmax to this distribution does not change the distribution, no matter the value of τ. Children then sample from their mental distribution to decide where to place their cushions; since there is only 1 bin with any probability, all samples yield that bin. So children are expected to stack their cushions in that bin, modulo α, the probability of placing a cushion at random. Importantly, the single simulation that children run will land in each possible bin in proportion to that bin's probability. So over many trials, these children will tend to stack in bins in proportion to their probability, modulo α. So if there are two possible bins and one bin is twice as likely as the other, they will stack in the higher-probability bin twice as often as they stack in the lower-probability bin.

Children who simulate once per cushion also generate a mental distribution with no variance, which is not impacted by the softmax function. They place a cushion in the simulated bin, modulo α, and repeat for each cushion. This process predicts that cushion placements will approximate the probabilities within each trial. So if there are two possible bins and one is twice as likely as the other, these children will

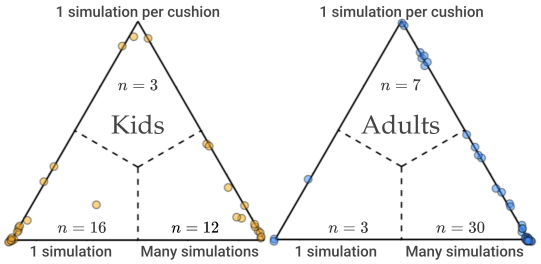

Figure 4: Model categorizations for children and adults. Points on a vertex most likely belong to the category on that vertex. Points on the perimeter (but not a vertex) are confusable between the categories on the adjacent vertices but not the category on the opposite vertex. Points on the perimeter at the midpoint between two vertices are maximally confusable between those two categories. Points in the interior of the triangle are confusable between all three categories, with points at the center of the triangle being maximally confusable between all three categories.

most likely put two cushions in the higher-probability bin and one cushion in the lower-probability bin (modulo α).

For each participant the model finds the probability that they (a) run 1 simulation (adults: mean probability = .09, 95% CI = [.02, .12]; 4-year-olds: mean probability = .47, 95% CI = [.30, .64]); (b) run 1 simulation per cushion (adults: mean probability = .23, 95% CI = [.09, .38]; 4-year-olds: mean probability = .19, 95% CI = [.04, .36]); or (c) run many simulations (adults: mean probability = .68, 95% CI = [.53, .83]; 4-year-olds: mean probability = .34, 95% CI = [.16, .50]).

Adult results validate the model: adults most likely run many simulations, and are unlikely to run a single simulation. The probability of an adult simulating once per cushion is not trivial (.23), but note that in this model simulating once per cushion is confusable with running many simulations with τ close to 1. Child results suggest that about half of 4-year-olds deploy minimal representations of possibility and that about half bring modal concepts to bear.

For each participant, the model gives a probability distribution over which of the three groups that participant belongs to. Fig 4 visualizes these distributions. Here we report how many participants most likely to belong to each group (1 simulation, 1 simulation per cushion, many simulations). Adult results validate the model: 30 adults (75%) were categorized as running many simulations; 7 (18%) as simulating once per cushion (note that simulating once per cushion is confusable with running many simulations with τ close to 1). Only 3 adults (8%) were categorized as running 1 simulation. Twelve children (39%) were categorized as running many simulations, 3 (10%) as simulating once per cushion, and 16 (52%) as running 1 simulation. Four-year-olds largely fall into two similar-sized groups: a group that runs a single simulation and a group who runs many simulations.

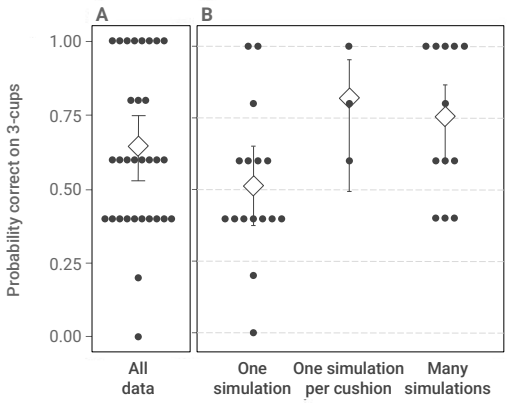

Figure 5: Performance on the 3-cups task. Large diamonds are model estimates with 95% CIs. Black dots are individual proportions of wise decisions. (A) All data. (B) Data broken down by model classifications.

3-cups task

We define a "wise decision" on the 3-cups task as placing the single cushion in the path of the ball that follows a determinate path, as this is the only way to guarantee a reward. Children placed their single cushion wisely 63% of the time; they put it in a merely possible bin 32% of the time. Performance is not significantly different from the 4-year-olds in Mody and Carey (2016) or Leahy (2023) study 2 (GLMM predicting probabililty of a wise decision from study with random intercept for participant: Leahy 2023-Current data: logOR -0.01, 95% CI [-0.68, 0.66], p = .98, Mody & Carey-Current data: logOR 0.41, 95% CI [-0.34, 1.16], p = .28). As in previous studies with different apparatuses, 4-year-olds make unwise decisions between 30% and 40% of the time. In addition to replicating previous results, we see here that 4-year-olds constrain their responses to bins that might pay off. This shows us that their errors draw on simulation, not noise. However, many children do not seem to differentiate bins that merely might pay off from bins that will pay off. Fig 5A shows the probability of a wise decisions, along with each participant's descriptive proportion of wise decisions. Note that the distribution of descriptive proportion correct is bimodal, with one mode at 100% correct and another mode centered on 50% correct. Distributions like this are typical on 3-cups tasks with children aged 3 or older; these distributions support the hypothesis that some children bring modal concepts to bear while others run a single simulation (Leahy et al., 2022; Leahy, 2023). The current data allows us to break the 3-cups data down my model categorization from the probability task (Fig 5B) to evaluate this claim.

Relationship between probability- and 3-cups tasks

We held out the 3-cups data in constructing our model of the probability task. Thus we can validate our model against the 3-cups data. We used the model categorizations to evaluate performance on the 3-cups task, checking whether children

who run a single simulation on the probability task choose the singleton in the 3-cups task with probability .5, and whether children who run many simulations in the probability task are more likely to make wise decisions on the 3-cups task. Each participant's descriptive proportion wise decisions, grouped by their model categorizations, are plotted in Fig 5B. The observed distribution of proportion correct for children who ran a single simulation on the probability task did not differ from the distribution expected if all children choose the singleton with ground-truth probability .5 (multinomial test, p=.31). This is a signature of minimal representations of possibility. Moreover, the probability of choosing wisely in the 3-cups task for these 16 children was not significantly different from existing studies with children aged 30-36 months, all of whom deploy minimal representations of possibility (GLMM predicting probability of a correct response on the 3-cups task from study (Mody & Carey 2.5-year-olds; Grigoroglou et al. 2.5-year-olds; Leahy 2023 2.5-year-olds, current 4-year-olds categorized as running a single simulation: all logOR between -0.25 and -0.08; all p-values between .44 and .78). The 4-year-olds who were categorized as running a single simulation on the probability task perform like 2.5-year-olds on the 3-cups task; all choose the singleton with ground-truth probability .5. By contrast, children who were categorized as running many simulations formed a distribution that was significantly different from the distribution expected if all children choose the singleton with ground-truth probability .5 (multinomial test, $p < .001$). To evaluate the differences between these groups a GLMM was fit, predicting the probability of a wise decision on the 3-cups task from model categorization (1 simulation, 1 simulation per cushion, or many simulations), with a random intercept for participant. Children who run many simulations on the probability task were significantly more likely to make a wise decision on the 3-cups task than children who run a single simulation (logOR 1.06, 95% CI [0.14, 1.99], p=.02). About half of 4-year-olds run a single simulation on both of these tasks, while the remainder run multiple simulations on the probability task and are more likely to differentiate the sure thing from the mere possibilities on the 3-cups task.

Discussion

The current study developed a dependent measure that is suitable for computational modeling. We found that on the probability task, 52% of 4-year-olds most likely run a single simulation; 39% most likely run many simulations; and 10% most likely simulate once per cushion (which is confusable with running many simulations with τ close to 1). This converges with existing estimates of the frequency of deploying minimal representations of possibility and deploying modal concepts among 4-year-olds (Leahy, 2023, 2024).

We replicated previous 3-cups findings. When 3-cups performance was broken down by model classifications from the probability task, we found that (1) children who were categorized as running a single simulation on the probability task performed indistinguishably from 2.5-year-olds on the 3-cups task and (2) children who were categorized as having a good sense of the probability distribution on the probability task were significantly better at the 3-cups task. This indicates that there is a common competence that both tasks measure, perhaps the ability to evaluate the results of multiple simulations of a single event.

The results shed light on the tension between infant success on probability tasks (Xu & Garcia, 2008) and preschooler struggles with the 3-cups task and Y-shaped tube task. Perhaps infants can evaluate probabilities as proportions in an observable population; correct evaluation of probabilities by repeated simulation emerges later.

Why would so many 4-year-olds run a single simulation? Two answers appear in existing literature. First, Leahy and Carey (2020) argue that children deploy minimal representations of possibility because they lack *modal concepts*: mental symbols that attach to a complete thought to mark that thought as merely possible, as one member of a range of competing simulated alternatives for a single reality. Lacking these symbols, children cannot incorporate multiple simulated possibilities into a single, coherent model; including two simulated outcomes ("The ball will land in the leftmost bin and the ball will land in the second bin from the left") results in an inconsistent, unuseable model. At most they can incorporate one simulated possibility into their model.

Second, running many simulations might take time or impose other costs. The greatest amount of information about a probability distribution is provided by the first sample from that distribution. So when sampling is costly enough, the most efficient policy is to sample once and take the result to be the fact of the matter (Vul, Goodman, Griffiths, & Tenenbaum, 2014). Our apparatus offers a way to test these hypotheses. We can force children to run multiple simulations by asking, "What will happen if the ball goes this way [indicating that the ball moves leftward]?" and "What will happen if the ball goes this way [indicating rightward]?" A child who correctly answers both of these questions has run multiple simulations. If some children are still categorized as running a single simulation with these prompts, we can conclude that they do not do so to save on simulation costs.

Conclusion

The current experiment had two key findings. First, our model lets us estimate that approximately 50% of 4-year-olds are best explained as running a single simulation per trial, while approximately 40% are best explained as running many simulations per trial. This converges with earlier findings that about half of 4-year-olds operate with a single simulation, and about half are responsive to multiple incompatible possibilities. Second, children who were categorized as running a single simulation on the probability task show a signature of minimal representation of possibility on the 3-cups task; this signature is not observed among children who run multiple simulations. There is something that both tasks measure. Perhaps both measure the ability to run multiple simulations.

Acknowledgments

This work was funded by NSF grant (2121009) awarded to J.B.T. and K.A.S. and by Navy-ONR grant N00014-23-1-2355 to J.B.T.

References

Alderete, S., & Xu, F. (2023). Three-year-old children's reasoning about possibilities. *Cognition*, *237*, 105472.

Andreuccioli, L., Mazor, S., Begus, K., Bonawitz, E., Denison, S., & Walker, C. M. (2024). Young children adapt their search behavior for necessary versus merely possible outcomes. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Brody, G., Mazalik, P., & Feiman, R. (2024). Object files encode possible object identities, but not possible locations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Cesana-Arlotti, N., & Halberda, J. (2024). A continuity in logical development: domain-general disjunctive inference by toddlers. *Open Mind*, *8*, 809–825.

Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature communications*, *11*(1), 5999.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Gerstenberg, T., Siegel, M., & Tenenbaum, J. (2018). What happened? In *Proceedings of the annual meeting of the cognitive science society.* PsyArXiv.

Leahy, B. (2023). Don't you see the possibilities? *Developmental Science*, *26*(6), e13400.

Leahy, B. (2024). Many preschoolers do not distinguish the possible from the impossible in a marble-catching task. *Joournal of Experimental Child Psychology*, *238*, 105794.

Leahy, B., & Carey, S. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78.

Leahy, B., Huemer, M., Steele, M., Alderete, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences*, *119*(52), e2207499119.

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48.

Redshaw, J., & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758–1762.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Turan-Küçük, E. N., & Kibbe, M. M. (2024). Three-year-olds' ability to plan for mutually exclusive future possibilities is limited primarily by their representations of possible plans, not possible events. *Cognition*, *244*, 105712.

Turan-Küçük, E. N., & Kibbe, M. M. (2025). Three-and four-year-old children represent mutually exclusive possible identities. *Journal of Experimental Child Psychology*, *249*, 106078.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2014). One and done? *Cognitive Science*, *38*(4), 599–637.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.