Cognition 266 (2026) 106283

Contents lists available at ScienceDirect

Cognition

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

Full length article

The development of probabilistic reasoning during early childhood

Sarah Placì [a,b,*], Stefania Pighin [a], Tommaso Mastropasqua [a], Katya Tentori [a]

[a] Center for Mind/Brain Sciences, University of Trento, Italy
[b] University of Göttingen, Germany

ARTICLE INFO

Dataset link: https://sarah-placi.github.io/prob_reasoning_preschoolers/index.html

Keywords:
Probabilistic reasoning
Proportions
Heuristic reasoning strategies
Preschoolers
Developmental psychology

ABSTRACT

Previous results have been inconsistent regarding the age at which children can compute probabilities based on proportions, with estimates ranging from one to 12 years. The aim of our study was twofold: (I) to address previous inconsistencies and (II) to quantify reasoning skills in 3- to 5-year-old children using an experimental procedure that is simple and engaging but also allows for the control of possible competing heuristic strategies, which may lead to outcomes mimicking the correct answer. Specifically, children had to choose between two urns containing varying proportions of blue and yellow balls. They knew they would receive a reward if, after several random mixtures, a target (e.g., a blue) ball fell out of the chosen urn. If children understood probabilities, they should select the urn with a higher proportion of target balls. Alternative reasoning strategies were disentangled by manipulating the absolute and relative numbers of target versus non-target balls in the two urns across 18 trials. We found that most 5-year-olds and a smaller but non-negligible number of younger children consistently chose the urn with a higher proportion of target items. However, a significant portion of 3- and 4-year-olds' responses appeared to be guided by heuristic strategies, with the most prevalent being choosing the urn with a higher number of target balls. These results provide a more comprehensive understanding of the development of probabilistic reasoning in young children and underscore the importance of using experimental procedures and stimuli that enhance reasoning abilities while controlling for competing strategies.

1. Introduction

Decades ago, Jean Piaget and Bärbel Inhelder wondered if humans have

> "... an intuition of probability just as fundamental and just as frequently used as, say, the intuition of whole numbers?" (Piaget & Inhelder, 1975, pp. xv).

After several years of research on children's reasoning abilities, the two researchers' verdict was that young children are devoid of probability intuitions. Instead, their findings suggested that the notions of chance and probability gradually appear around age 7, parallel with the differentiation between reversible operations that lead to deductible outcomes and irreversible operations that induce uncertainty. Only around age 12, Piaget and Inhelder concluded, do children start understanding how to characterize uncertainty using probabilities. Despite the thoroughness of their investigation, Piaget and Inhelder's methods faced criticism. One potential problem was their over-reliance on children's verbal skills and explicit judgments (see, for example, Yost, Siegel, & Andrews, 1962), which might not be aligned with their implicit reasoning. Nevertheless, more than half a century later, there still is no satisfying answer as to when probabilistic reasoning develops.

One way to evaluate young children's understanding of probabilities is to examine their choices when faced with random draws from unbalanced urns. Imagine, for example, you are shown two urns: A, which contains 8 blue and 12 yellow balls, and B, which contains 12 blue and 8 yellow balls (see Example 1 in Fig. 1). Your task is to choose a urn, knowing that you win a prize if a target ball is randomly drawn from it. The normative way to quantify the probability of randomly drawing a target ball from each urn is to estimate the proportion between the number of possible ways in which the wished event (i.e., obtaining a blue ball) could happen and the total number of possible outcomes. In our example, therefore, the probability of randomly drawing a blue ball from A is $\frac{8}{8+12} = 0.4$. The probability of randomly drawing a blue ball from B is $\frac{12}{8+12} = 0.6$. If your goal is to maximize your likelihood of winning a prize, the math we just did should convince you to pick B.

However, in this precise case—as in many other instances in psychological research—a correct response does not necessarily warrant the endowment of the system of formal thought under investigation. This

* Correspondence to: Georg-Elias-Müller-Institute of Psychology, Gosslerstrasse 14, 37073 Göttingen, Germany.
E-mail address: sarah.placi@uni-goettingen.de (S. Placì).

https://doi.org/10.1016/j.cognition.2025.106283
Received 1 October 2024; Received in revised form 13 January 2025; Accepted 30 July 2025
Available online 22 August 2025

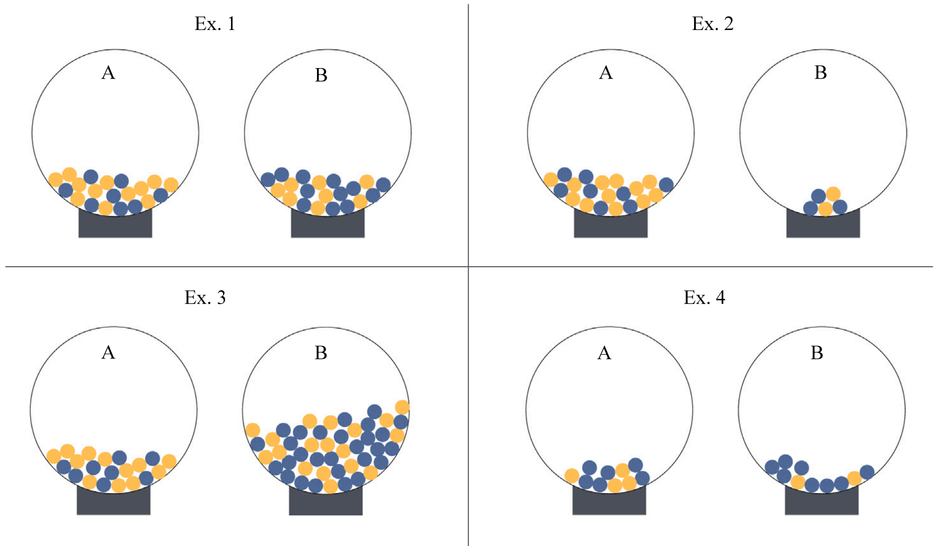

Fig. 1. Examples of containers used to investigate probabilistic reasoning.
In the examples above, the target color is blue, and the favorable urn is always B (although, of course, during the experiments, the target color and the position of the favorable urn are typically counterbalanced). In Example 1, a correct answer could be achieved by using several heuristic reasoning strategies. In Example 2, choosing the container with more target items (i.e., using the 'More good' strategy) would lead to the wrong answer. In Example 3, picking the container with fewer non-target items (i.e., using the 'Less good' strategy) would lead to the wrong answer. In Example 4, choosing any container having more target than non target items (i.e., using the 'Good urn' strategy) would result in a random choice. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is because there are alternative reasoning strategies that could have led to the belief that B is the best choice. In particular, you could have used a simplified heuristic in the form of "the more target balls in one urn, the higher the chance of getting one of them out of it". In our example, the outcome of this strategy would align with the correct answer but could not be qualified as rational (and indeed it would only lead to success in specific situations).

Imagine now a different pair of urns: A containing 8 target balls and 12 non-target balls, and B containing 3 target balls and 2 non-target balls (see Example 2 in Fig. 1). By the above-mentioned heuristic, you would choose A, which is the wrong answer since A contains a greater absolute number but a lower proportion of target balls than B (i.e., 8 vs. 3, but 0.4 vs. 0.6, respectively). On the other hand, if you provide the correct answer (i.e., choosing B), we could be rather confident that your reasoning is not guided by the simplified strategy at issue. However, yet again, we could not exclude that such a response is driven by another heuristic strategy (see below for more examples of alternative heuristic strategies), making it all but trivial to conclude whether you have a proper understanding of probabilities.

Different versions of the two-urn paradigm have been used to test whether young children quantify probabilities based on proportions (see, for example, Denison & Xu, 2010, 2014; Falk, Yudilevich-Assouline, & Elstein, 2012; Girotto, Fontanari, Gonzalez, Vallortigara, & Blaye, 2016; Goldberg, 1966; Gualtieri & Denison, 2019; Placì, Fischer, & Rakoczy, 2020; Yost et al., 1962). The advantage of this paradigm is that it does not necessarily require children to speak to express their decision. They can point or move toward their chosen urn as well. To assess children's reasoning strategies (i.e., whether they rely on proportions to quantify probabilities or, instead, on different heuristic reasoning strategies), they must typically undergo several trials in which absolute and relative quantities are disentangled (see, for example, Denison & Xu, 2014; Falk et al., 2012; Girotto et al., 2016; Gualtieri & Denison, 2019; Placì et al., 2020, who also controlled for several heuristic reasoning strategies). For example, as seen above, children who base their choice on the absolute number of target objects (which we will call the 'More good' strategy) should perform poorly when the number of target objects is higher in the urn with a lower proportion of target objects (as in Example 2 of Fig. 1). In other conditions, what might be misleading is the number of non-target objects, that is higher in the urn with a higher proportion of target objects (see Example 3 of Fig. 1). Children who base their decision on picking the urn with fewer non-target objects (which we will call the 'Less bad' strategy) should perform poorly in these conditions. Another strategy children might use is simply choosing any urn with more target than non-target objects (which we will call the 'Good urn' strategy). If this is the case, whenever both urns have more target than non-target objects (see Example 4 of Fig. 1) or both urns have less target than non-target objects, children's performance would be at chance level, as they would face a random choice between two 'good' or two 'bad' urns.

It is worth noting that the two-urn paradigm does not require mastery of math. Children can provide the correct answer by visually comparing the proportions of items, without counting or making any calculation, an ability that 6-month-old babies already seem to possess (McCrink & Wynn, 2007). Furthermore, it is important to emphasize that the ability to visually compare proportions is a prerequisite in this paradigm, but it is not the focus of the test. What is being scrutinized here is the understanding that a higher proportion of target items corresponds to a higher probability of obtaining a target item.

Among the first studies investigating young children's ability to quantify probabilities based on proportions using a two-urn paradigm were a study by Yost et al. (1962) with 4- and 5-year-olds and a study by Goldberg (1966) with 3- to 5-year-olds. In both studies, two urns were used, each containing chips of the children's preferred color (as established during the experimental procedure) in equal numbers but in different proportions relative to the number of chips of another color. Preschoolers had to choose the urn from which to draw a target chip (which was of the preferred color in half of the trials and of the other color in the remaining half). Although the two studies used very similar methods, their conclusions differed: While according to Yost et al. (1962) the children they tested relied on the proportions of target items to make their decisions, Goldberg (1966) argued that most of their participants relied on comparing the absolute number of target items (i.e, used the 'More good' strategy), because their performance was worse in those trials in which the number of target items was equally distributed in the two urns.

More recently, Falk et al. (2012) tested 4- to 11-year-old children, using a comprehensive set of trials to control for reliance on various quantity heuristics, several of which were not controlled before.

Their findings suggested that children start relying on proportions to compute probabilities at around seven. Before that age, Falk et al. (2012)'s results indicated that children primarily rely on the number of target objects (i.e., the 'More good' strategy) to evaluate chances. Their findings were, therefore, coherent with those of Goldberg (1966).

What came to challenge the 'pessimistic' conclusion above was work conducted with much younger children, specifically 12-month-old infants (Denison & Xu, 2010, 2014). In two studies that completed each other (i.e., the second controlling for the heuristics that the first did not control for), infants were presented with two urns containing a more and a less preferred kinds of lollipops, respectively (determined beforehand with a preference trial). The infants had to choose between two opaque cups, which concealed a lollipop that the experimenter had secretly drawn from one of the two urns. The authors examined whether the infants preferred (by physically approaching it) the cup containing the lollipop drawn from the urn with the higher proportion of their favorite kind of lollipop. Several conditions were manipulated between participants to control their reliance on different heuristic reasoning strategies. The results, this time, were strikingly different than those of previous research: in all the conditions, most infants chose the cup containing the item drawn from the urn with the higher proportion of target items. The authors, thus, concluded that probabilistic inferences are in the cognitive repertoire of infants[1] and went back to Laplace's famous statement that "The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which times they are unable to account" (Laplace, 1814).

Girotto et al. (2016) conducted a close replication with 3-year-olds of the Denison and Xu (2014)'s study. In addition, they also tested 3- to 5-year-olds in a modified version of that paradigm where red and black chips were used as a target and non-target items and where children were given a reminder representing the proportions of chips before the drawing. They found that 5-year-olds performed generally well, i.e., they chose the cup hiding the objects drawn from the containers with a higher proportion of target objects. On the other hand, 3-year-olds performed at chance level in almost all trials, and 4-year-olds performed at chance level in trials in which quantity and proportions were dissociated (i.e., when the urn with the more favorable proportion contained the lower number of target items). In light of these results, the authors concluded that children's verbal judgments, as well as their explicit choices, did not reveal any ability for probabilistic inferences before the age of 5. They put into question the development of probabilistic reasoning in young children, wondering whether it followed a U-shaped curve or whether infants' ability had been overestimated.

A further study (Placì et al., 2020) directly compared the performance of 12-month-olds and 3- and 4-year-olds in a modified version of the Denison and Xu (2014)'s study. In this case, children were only tested in a control trial with deterministic outcomes (i.e., one urn only had the favorite kind of objects and the other urn only the less favorite kind) and in a trial in which the absolute number of favorite items was misleading (i.e., lower in the urns with a higher proportion of favorite items). Both the infants and the preschoolers failed in the probabilistic task, and only the preschoolers performed above chance level in the deterministic task. These results are in line with findings that suggest a lack of probability intuitions in preschoolers (Falk et al., 2012; Girotto et al., 2016; Goldberg, 1966) while contrast with those indicating that infants are proficient probabilistic reasoners (Denison & Xu, 2014).

Finally, some years later, Gualtieri and Denison (2019) conducted a study with 3- and 4-year-olds using virtual urns. Here too, children underwent several trials in which relative and absolute quantities of target items were dissociated in various ways. In contrast to previous studies conducted with preschoolers, however, the participants chose the urn with the higher proportion of target items more often than expected by chance in all trials. To our knowledge, this was the first evidence suggesting that 3-year-olds are capable of this kind of probabilistic reasoning.

Considering the contrasting findings reported above, there is still work to do before we can confidently answer the question raised by Piaget and Inhelder decades ago. It is our hope, however, that by considering the main factors that may have influenced the results presented in the previous paragraphs, we can better understand their heterogeneity and devise an improved way of investigating young children's probabilistic reasoning skills. For this reason, we pursued the following two goals: (i) to address the possible limitations of previous studies in order to identify the factors affecting children's performance, and (ii) to test 3-, 4-, and 5-year-olds in a novel experiment that employed a procedure and stimuli which, although simple and engaging, allowed us to better interpret children's performance.

2. Addressing previous inconsistencies

The studies discussed above all employ a two-urns paradigm to investigate probabilistic reasoning in children. However, they differ by several factors that are worth considering as these might explain some of the variance found in their findings. In what follows, we have organized these factors into three sections, based on whether they concern the experimental procedure, the material (i.e., the stimuli), or the analyses.

2.1. Different procedures

2.1.1. Demand on executive functions

One first procedural difference between the infant study (Denison & Xu, 2014) and the failed replication with preschoolers (Girotto et al., 2016) is that the infants were being held by their parents before choosing a cup and could, therefore, concentrate on the information being presented, while the preschoolers had to regulate their spontaneous behavior on their own, which might have made it more difficult for them to concentrate on the task at hand. In other words, more demands could have been put on preschoolers' executive functions (e.g., inhibitory control) than on the infants'. One way to reduce the demands is to shorten the procedure to a minimum length to reduce the time during which children have to inhibit their spontaneous behavior. In line with that, Gualtieri and Denison (2019), who found a better performance from 3- and 4-year-olds (in comparison to Girotto et al. (2016)), used a simplified version of the two-urn paradigm. Specifically, they used virtual urns displayed on a screen: A program handled the content of the urns (i.e., shuffling the balls and drawing them), which reduced the time needed by the experimenter to handle the urns and mix and draw the items from them.

2.1.2. Inadvertent parental interference

Infants' remarkable performance reported in Denison and Xu (2014) could also have been influenced by another procedural aspect: the infants were held on the ground in a 'neutral' position by their parents before being allowed to crawl toward their preferred sample. It cannot be ruled out that, when instructed to release their children, some of these parents inadvertently directed them toward the 'correct' sample. In line with this possibility, Placì et al. (2020) attempted to replicate this study while asking parents to wear opaque glasses. In this case, infants were found to perform at chance level not only when the

[1] Note that this work with infants inscribed itself in a series of other studies testing statistical reasoning in infants (Denison, Reed, & Xu, 2013; Denison & Xu, 2019; Xu & Denison, 2009; Xu & Garcia, 2008; Xu & Kushnir, 2013). It did, thus, not stand alone in suggesting that infants are good statistical reasoners, nor that they are capable of probabilistic inferences about a limited number of events (Téglás & Bonatti, 2016; Téglás, Girotto, Gonzalez, & Bonatti, 2007; Téglás, Ibanez-Lillo, Costa, & Bonatti, 2015; Téglás, Vul, Girotto, Gonzalez, Tenenbaum, & Bonatti, 2011). However, this study was (and still is to our knowledge) the first of its kind indicating that infants can explicitly quantify probabilities based on proportions.

absolute number of targets and proportion were dissociated but even in a simpler condition with only deterministic outcomes (i.e., one of the two urns contained only target objects and the other urn only non-target objects).

2.1.3. Epistemic versus aleatory uncertainty

Another difference across studies was that, in some cases, participants had to choose between *objects that had already been drawn* from the urns, whereas, in others, they had to choose between *urns from which to draw an object*. Although from a strictly normative point of view, there is no difference between these two cases, there is evidence that reasoning about probabilities is easier for children when uncertainty expresses inherent variability of a physical environment (Robinson, Rowley, Beck, Carroll, & Apperly, 2006). In Girotto et al. (2016), the uncertainty was epistemic, meaning the objects had already been drawn, but participants lacked precise information about them; in Gualtieri and Denison (2019), uncertainty was aleatory and resided in the environment, as the drawing had not yet occurred. This could (at least partly) explain why preschoolers performed better in the latter than in the former study.

2.2. Different stimuli

2.2.1. Likelihood ratio

Another factor that could have impacted children's performance concerns the difficulty in discriminating the probability of yielding a target object from each of the two urns (i.e., the likelihood ratio between them, hereafter 'LR'). For example, in their non-deterministic trials (i.e., trials in which neither urns had a proportion of targets equal to 0 or 1), Denison and Xu (2014), Girotto et al. (2016), Goldberg (1966), Gualtieri and Denison (2019), Placì et al. (2020) and Yost et al. (1962) employed LRs ≥ 2, meaning that the proportion of targets was at least twice as high in one urn compared to the other. In contrast, in almost all their trials, Falk et al. (2012) used LRs ≤ 2. For example, in Trial 21 (see Table 3 of Falk et al., 2012), there were 3 target objects versus one non-target object in one urn (i.e., a 0.75 proportion), and 5 target objects and 2 non-target objects in the other urn (i.e., a 0.71 proportion), resulting in a LR of 1.05. Such small LRs might be much harder to discriminate, and raise the issue whether children in this study primarily relied on heuristic reasoning strategies because they lacked any concept of probabilities or because the answer would have necessitated actual counting.

2.2.2. Two versus three dimensional stimuli

Another reason participants could have performed better in some studies than in others could be due to the dimensionality of the object to draw. Specifically, in the study in which preschoolers' performance was the best (i.e., Gualtieri and Denison (2019)), children were presented with two-dimensional urns filled with two-dimensional balls. In the other studies, children were presented with actual urns filled with real objects. Children might find it easier to extract statistical information from urns filled with two-dimensional balls that have identical features (i.e., that are made of a similar composition of pixels) than from urns filled with three-dimensional objects, which, depending on the viewing angle and obstruction by other objects, may not be as easily recognized or quantified.

2.2.3. Overlapping between responses based on normative and alternative strategies

An important limitation of some previous studies, which could have led to an overestimation of children's performance, was the lack of consideration for the possible overlap between normative responses and responses generated by alternative strategies. This means that some good performances might have been the results of non-normative strategies rather than an expression of sound probabilistic reasoning. For example, in the trials in which Gualtieri and Denison (2019) and Girotto et al. (2016) aimed to control whether children relied on the 'Good urn' strategy by choosing any urn in which there were more target than non-target objects, (see trial Type A in Gualtieri and Denison (2019) and Task B2 in Girotto et al. (2016)), one of the urns was deterministic (i.e., it contained only target objects). This means that this urn was different not only from a *probabilistic* point of view but also from a *modal* perspective, as drawing from it would *necessarily* yield a target object. Similarly, in the trials aimed at controlling for the 'Less bad' strategy by choosing the urn with fewer non-target objects in Gualtieri and Denison (2019), one of the urn contained only non-target objects (see trial type E in Gualtieri & Denison, 2019). Given this, children could have solved these trials by relying on modal concepts, without quantifying probabilities.[2] In trial Type A of Gualtieri and Denison (2019) and Task B2 of Girotto et al. (2016), children could have chosen the *necessary* outcome over the *possible* outcome, and in trial type E of Gualtieri and Denison (2019), they could have chosen the *possible* outcome over the *impossible* one. Moreover, in all remaining trials, children could have succeeded by choosing the urn with more targets than non-target objects, given that the 'Good urn' strategy aligned with the normative response in these trials.

2.3. Different analyses

Another limitation of some previous studies is that they did not take important factors into account in their analyses and only reported results at the group level. For example, Goldberg (1966) tested 3- to 5-year-olds but only reported overall results, without distinguishing them by age. Their conclusion that children primarily relied on the 'More good' strategy could have then been driven mainly by the poorer performance of the younger children. Yost et al. (1962), who used similar methods but tested only 4- and 5-year-olds, reported better overall results. However, Yost et al. (1962) did not check whether children might have relied on the 'More good' strategy, even though they constructed their trials in a way that allowed such an analysis. Goldberg (1966) checked whether children's performance was lower in the trials in which the 'More good' strategy would have been inconclusive and found that it was the case. These different analyses could explain the difference in the findings reported by both studies.

While in some studies individual-level analyses were omitted simply because there were too few data points per participant (see, for example, Denison & Xu, 2014; Placì et al., 2020), other studies with enough data points per participant still failed to report on individual-level performances, or did so only partially (see, for example, Girotto et al., 2016; Gualtieri & Denison, 2019; Yost et al., 1962). Group-level analyses can reveal the trial types that children tend to find more challenging, indirectly informing us about the strategies they were likely using. For example, observing chance performance in trials where choosing the urn with more target objects should lead to incorrect choices would suggest that at least some children relied on the 'More good' strategy. However, an analysis that also looks at the individual performance could, in addition, provide more data on how many children at each age rely on each specific strategy (see, for example, Falk et al., 2012, who analyzed results at the individual level). Such information would help us better track the development of probabilistic reasoning because it tells us at what age the ability first appears and when it reaches full maturity.

[2] Note that the ability of preschoolers to reason about possibilities to solve problems is also debated (see, for example, Leahy, 2023). However, recent findings suggest that when presented with visual displays of objects from which items are randomly drawn, 3-year-olds can use modal concepts (Alderete & Xu, 2023; Stahl & Feigenson, 2024).

3. Experiment

The present study aims to provide a clearer picture of the developmental trajectory of probabilistic reasoning in preschoolers. In light of the limitations and issues identified in previous studies, the version of the well-established two-urn paradigm used in the present experiment incorporated the following features. First, to rule out the possibility that participants might respond by applying modal concepts, we did not use deterministic urns in any of the test trials (i.e., in all test trials, the proportion of target balls in each urn was never equal to 1 or 0). Second, to avoid possible difficulties with real-world 3-D objects, that can be partly or wholly concealed, a virtual (animated) 2-D version of the urns was employed. Third, given the above-mentioned finding that aleatory uncertainty might be more easily understood than epistemic uncertainty, we asked participants to choose which urn to draw an object from, rather than choosing between objects that had already been drawn. Fourth, as previous studies found different results with different LRs, we explicitly manipulated this variable, by using an 'easier' LR (= 4) and a 'more difficult' LR (= 1.5). Fifth, and most importantly, our experiment involved multiple trials to control for the use of various heuristics and to allow for both group-level and individual-level performance analyses.

An illustration of what the task looked like is provided in Fig. 2. Children won a sticker every time they chose the urn that produced a ball of the target color. The target color was kept constant for each child but was counterbalanced between children (i.e., it was blue for half of them and yellow for the other half). All participants underwent 4 familiarization trials (in half of these the experimenter selected the urn while in the other half the child was asked to do so), 2 control trials (in which one urn contained only target objects and the other only non-target objects), and 16 test trials (see Table 1 for a detailed description of the 22 trials).

The test trials were designed to assess whether participants relied on proportions to estimate the probability of drawing a target ball or used other strategies. Specifically, we considered the following strategies: (i) whether participants engaged in proper probabilistic reasoning by comparing proportions of target balls in the two urns (the normative 'Ratio' strategy), (ii) whether participants answered randomly (the 'Random' strategy), (iii) whether participants compared the urns based on their quality, with 'good' urns having more target than non-target balls and 'bad' urns having more non-target than target balls (i.e., the 'Good urn' strategy), (iv) whether participants chose the urn with the greater absolute number of target balls (i.e., the 'More good' strategy), (v) whether participants chose the urn with the lower absolute number of non-target balls (i.e., the 'Less bad' strategy).

3.1. Methods

3.1.1. Ethical statement

The experiment was approved by the ethics committee of the University of Trento for experimentation with human beings (protocol number 2020-003).

3.1.2. Participants

The data were collected between May 2021 and November 2022. Participants were recruited through the Trento province (Italy) public schools and word of mouth. The sample consisted of 90 children (45 girls), divided into three age groups: 3-year-olds ($n = 30$; $M_{age} = 3.66$; age range = 3 years and 1 month – 3 years and 11 months), 4-year-olds ($n = 30$; $M_{age} = 4.57$; age range = 4 years – 4 years and 11 months) and 5-year-olds ($n = 30$; $M_{age} = 5.63$; age range = 5 years – 5 years and 11 months). Seven additional children were excluded for the following reasons: 3 due to parental interventions providing feedback (despite initial agreement to refrain from doing so); 2 due to the children's unwillingness to complete the experimental session; and one due to an error made by the experimenter. All children were Italian native speakers and attended public preschools. Parents signed an informed consent before their child participated in the experiment.

3.1.3. Design, materials, and procedure

The experiment used a 3 (Age: 3 vs. 4 vs. 5; between) × 2 (urn size: small vs. large; within) × 2 (LR: 4 vs. 1.5; within) × 4 (trial type: A vs. B vs. C vs. D; within) mixed design. The different factors will be explained in more detail below.

A demo version of the study can be run on the repository site here https://sarah-placi.github.io/prob_reasoning_preschoolers/index.html.

The study was programmed to run in a web browser using HTML and JavaScript. To give participants a realistic impression of the objects moving around in the two urns, the animation was programmed using the physics simulator Box2dweb (https://github.com/hecht-software/box2dweb), the JavaScript adaptation of Box2D (https://box2d.org/). The experiment was created as an interactive video game to increase participants' attention and their engagement in the task. For example, some of the game elements we introduced were the 'Clapping' and 'Sad Trombone' sounds, which played after a target or a non-target ball fell out of the selected urn, respectively.

Participants were tested individually via an online video conferencing platform (e.g., Skype or Zoom) in the presence of one parent. Before the start of the experiment, parents had to confirm that they would not intervene (e.g., by giving their child feedback or making comments about the task) unless it was to assist the experimenter, at her explicit request, in confirming their child's responses (e.g., if the child was pointing at the urns rather than verbally expressing their choice).

Each experimental session was recorded with the consent of the parents. The experimenter shared her screen and started the experiment on her computer. In the beginning, the experimenter gave a short introduction and explanation of the 'game'. Children first saw two balls, one blue and one yellow, and were asked to name their colors. This ensured that they could correctly perceive and identify the colors used in the experiment. On the next screen, the children saw two urns displayed next to each other. The experimenter told the participants there were "a number of blue and yellow balls" in each urn. She pointed out to participants that the rules and procedures of the game would always be the same. First, the balls were mixed. Then, the player had to choose one of the two urns (the experiment chose the left urn to give an example). The balls were mixed a second time, and, while they were being mixed a third time, a black mask rose and hid the content of the urns. A ball would finally fall from the chosen urn. The experimenter emphasized that the ball only fell from the selected urn. After that, the experimenter showed a second example of the procedure, this time choosing the urn on the right. These two examples constituted the first familiarization trials. Once the second example was completed, the experimenter told the participants they could start playing and that, to win a sticker,[3] a blue (for half participants) or a yellow (for the other half) ball had to fall from the urn they chose. The following two trials were still considered familiarization trials (though participants were unaware of it). The proportion of target balls was 0.5 in all four familiarization trials, meaning there was no correct answer (i.e., neither urn was preferable over the other in terms of the probability of yielding a target ball). In this way, by not providing participants with any probabilistically motivated outcome, we limited potential learning during this phase.

To prevent memory issues, at the beginning of each trial, the experimenter reminded participants of the target color. She also instructed them to observe the balls in the urns before mixing them. Only then, she asked participants to choose the urn from which they wanted the ball to be drawn. Children could respond either verbally or by pointing to one of the urns.

Whenever a ball of the target color was drawn, a happy face appeared between the urns, accompanied by a clapping sound. The experimenter also said the following phrase (translated from Italian): "Good,

[3] The stickers had previously been sent to children's home and during the game were handed out by their parents.

Table 1
Description of the urns.

	Trial number	Trial type	Urn 1			Urn 2			LR
			Target balls	Non-target balls	Proportion of target balls	Target balls	Non-target balls	Proportion of target balls	
Familiarization	1		4	4	**0.5**	4	4	**0.5**	1
	2		8	8	**0.5**	8	8	**0.5**	1
	3		3	3	**0.5**	3	3	**0.5**	1
	4		6	6	**0.5**	6	6	**0.5**	1
Control	1		8	0	**1**	0	8	**0**	Infinite
	2		0	16	**0**	16	0	**1**	Infinite
Test	1	A	8	2	**0.8**	2	8	**0.2**	4
	2	A	4	16	**0.2**	16	4	**0.8**	4
	3	B	6	9	**0.4**	3	27	**0.1**	4
	4	B	6	54	**0.1**	12	18	**0.4**	4
	5	C	2	3	**0.4**	3	27	**0.1**	4
	6	C	6	54	**0.1**	4	6	**0.4**	4
	7	D	10	15	**0.4**	1	9	**0.1**	4
	8	D	2	18	**0.1**	20	30	**0.4**	4
	9	A	6	4	**0.6**	4	6	**0.4**	1.5
	10	A	8	12	**0.4**	12	8	**0.6**	1.5
	11	B	9	1	**0.9**	3	2	**0.6**	1.5
	12	B	6	4	**0.6**	18	2	**0.9**	1.5
	13	C	9	1	**0.9**	15	10	**0.6**	1.5
	14	C	30	20	**0.6**	18	2	**0.9**	1.5
	15	D	27	3	**0.9**	3	2	**0.6**	1.5
	16	D	6	4	**0.6**	54	6	**0.9**	1.5

a blue/yellow ball fell out". Conversely, if a ball of the non-target color was drawn, a sad face appeared between the urns along with a grave trombone sound. In this case, the experimenter said the following phrase (again translated from Italian): "Too bad, a blue/yellow ball fell out, but do not worry, there are other trials!"

To make the task more realistic and to avoid interfering with the children's genuine probability intuitions, all draws were fair, meaning that the outcome of each trial was randomly determined by the program based on the underlying proportion of the balls of the two colors in the selected urn.

The main part of the experiment consisted of 18 trials: 16 test trials and 2 control trials. Only these 18 trials were included in the analyses. Their presentation order was random, while the side of the 'more favorable' urn (i.e., the urn containing the higher proportion of target balls) was counterbalanced across trials. The control trials were added to ensure that participants understood the task properly and were engaged with it. In these trials, one urn contained only target balls (i.e., the proportion of target balls was 1), while the other urn contained only non-target balls (i.e., the proportion of target balls was 0). Participants did not need a concept of probability to answer correctly in these trials, but they had to understand how the task worked, know which color was the target, and be able to either voice their choice or to point to it.

The complete list of the trials (4 familiarization, 2 control, and 16 test), including the numbers of target balls in each urn, the corresponding proportions, and LRs, is provided in Table 1. Note that half of the trials are 'replications' of the other half, with the number of balls simply doubled. Take for example test trials 1 and 2. In trial 1, the favorable urn is Urn 1 because it has a higher proportion of target balls (i.e., 0.8). Specifically, it contains 8 target and 2 non-target balls. In test trial 2, the favorable urn is Urn 2, which contains exactly double the amount of target and non-target balls than Urn 1 (16 and 4 respectively). The same is true for the less favorable urns. This allowed us to present the relevant trials twice, by keeping the proportions constant without using exactly the same urns. In the design and analyses, we will refer to this factor (i.e., the difference in the absolute number of balls) as 'urn size'.

The 16 test trials instantiated two different LRs: 4 and 1.5. They can also be grouped into four different trial types, which we labeled (both in Tables 1 and 2): A, B, C, and D. These different trial types code the kind of reasoning strategies that lead to the choice of the correct urn (i.e., the one with the higher proportion of target balls). A graphical illustration of the different trial types is given in Fig. 2. Table 2 shows the trial types and the urn choice according to each strategy.

As can be extrapolated from Table 2, the different strategies produce varying rates of correct responses. Of course, the normative 'Ratio' strategy always leads to the choice of the correct urn, and therefore consistent use of this strategy would result in a 100% accuracy rate (i.e., 16 correct judgments out of 16). Using either the 'More good' or 'Less bad' strategies would result in an accuracy rate of 75% (i.e., 12 correct judgments out of 16). Indeed, the 'More good' strategy leads to the choice of the wrong urn only in type C trials while the 'Less bad' strategy does so only in type D trials. With regard to the 'Random' strategy, the accuracy rate can only be 'expected' and, since there are only two possible responses, it should average around 50% (i.e., 8 correct judgments out of 16), independently of the type of trial. The 'Good urn' strategy leads to the correct answer in type A trials, while is expected to lead to the choice of the correct urn only in half of the other cases (i.e., in type B, C and D trials). Therefore, the rate of correct responses produced by this strategy is deterministic for 4 judgments and random for the remaining 12 judgments, and should average around 63% (i.e., 10 judgments out of 16).

Therefore, type A trials are the only ones that can be answered correctly by applying any strategy other than the random one. In type B trials, all strategies led to the correct answer, except for the 'Random' and the 'Good urn' strategies, which were expected to lead to the correct answer only in half of the cases. Type C trials are the only trials that cannot be answered correctly using the 'More good' strategy, while type D trials are the only trials that cannot be answered correctly using the 'Less bad' strategy. Take the examples illustrated in Fig. 2. Imagine a participant relying solely on the 'More good' strategy. In our example, they should select the left urn in all trials except for type C trials. In these trials, they should select the wrong urn because there are more target objects in the unfavorable urn (i.e., the urn on the right side in Fig. 2). On the other hand, if the participant performs well in all trials but type D ones, this would suggest that they rely on the 'Less bad' strategy, because in these trials the urn with fewer non-target items is the less favorable one. Finally, if a participant performs well especially in Type A trials, this would suggest that they rely on the 'Good urn' strategy, because these are the only trials in which there is a 'good' urn (i.e., an urn that contains more target than non-target objects; this is the left urn in Fig. 2) and a 'bad' urn (i.e., an urn that contains more non-target than target objects; this is the right urn in Fig. 2). Trivially, if

Table 2
Urn choice according to each strategy in the 16 test trials. Bold numbers are trials in which a given strategy does not (always) lead to a correct answer.

Trial number	Trial type	LR	Correct urn	Predicted urn by each strategy				
				Ratio	Random	Good urn	More good	Less bad
1	A	4	1	1	**1|2**	1	1	1
2	A	4	2	2	**1|2**	2	2	2
3	B	4	1	1	**1|2**	**1|2**	1	1
4	B	4	2	2	**1|2**	**1|2**	2	2
5	C	4	1	1	**1|2**	**1|2**	**2**	1
6	C	4	2	2	**1|2**	**1|2**	**1**	2
7	D	4	1	1	**1|2**	**1|2**	1	**2**
8	D	4	2	2	**1|2**	**1|2**	2	**1**
9	A	1.5	1	1	**1|2**	1	1	1
10	A	1.5	2	2	**1|2**	2	2	2
11	B	1.5	1	1	**1|2**	**1|2**	1	1
12	B	1.5	2	2	**1|2**	**1|2**	2	2
13	C	1.5	1	1	**1|2**	**1|2**	**2**	1
14	C	1.5	2	2	**1|2**	**1|2**	**1**	2
15	D	1.5	1	1	**1|2**	**1|2**	1	**2**
16	D	1.5	2	2	**1|2**	**1|2**	2	**1**

participants performed well only in half of the trials, this would indicate that they choose randomly (i.e., the 'Random' strategy).

Importantly, the different quantity comparisons required by the different strategies, whether absolute or relative, never differed by a ratio smaller than 2:3, meaning that one quantity was always at least 1.5 times bigger than the other. Since children can discriminate visual quantities that differ by a 2:3 ratio at ten months (Xu & Arriaga, 2007), this ensured us that our participants could use each of the strategies we set out to assess. Take for example test trial 5 (see Table 1). Relying on the 'Good urn' strategy requires children the ability to compare the number of target and non-target balls within the same urn. In test trial 5, there were 2 target and 3 non-target balls in Urn 1. These two numbers differ by a ratio of exactly 2:3, meaning that children should have been able to perceive this difference and considered Urn 1 as a "bad" urn. In Urn 2, there were 3 target and 27 non-target balls, i.e., 9 times more non-target than target balls, which should have been an even easier comparison to make. Relying on the 'More good' strategy requires children to be able to compare the number of target balls between both urns, which in trial 5 differed again by a ratio of 2:3 (i.e., 2 target balls in Urn 1 and 3 in Urn 2). Finally, relying on the 'Less bad' strategy requires children the ability to compare the numbers of non-target objects, which here differed by a ratio of 1:9 again.

3.2. Predictions

We predicted an ordinal age effect: older children should perform better than younger children. Previous studies that tested children within a similar age range have observed this effect (Falk et al., 2012; Girotto et al., 2016; Gualtieri & Denison, 2019). We did not predict an effect of urn size, as this factor was not shown to matter in previous studies (see, for example, Falk et al., 2012; Goldberg, 1966). Given the inconsistencies in previous studies, we did not have specific predictions concerning trial type and LR. However, we assumed that LR would affect participants who can compare the proportions of target balls. Specifically, being confronted with a smaller LR should make the task harder and, consequently, lead to decreased performance. Trial type, on the other hand, should only affect those participants who rely on quantity heuristics. Previous studies indicated preschoolers' most prevalent heuristic strategy is the 'More good' strategy. For this reason, the worst performance was expected to occur in type C trials, that is, the trials for which this strategy leads to the choice of the wrong urn. We also had no specific predictions for the effect of trial order (e.g., whether performance improves during the experiment). However, we considered it in our analysis to account for a potential learning effect.

3.3. Results and discussion

The data sets and the analysis script are provided on the repository site. All analyses were conducted in R (R Core Team, 2021). We first analyzed group-level performance and then also looked at individual-level performance.

3.3.1. Group-level performance

As an initial step, we examined whether participants from the different age groups performed well in the 2 control trials. We did so by computing the proportions of correct answers in these trials and contrasting them with chance levels by running one-sample proportions tests with continuity correction, using the prop.test function from R's stats package. We found that the proportion of correct answers in the control trials was significantly above chance for all age groups: 0.92, 95% CI $[0.81, 0.97]$, $p < 0.001$, for the 3-year-olds; 0.85, 95% CI $[0.73, 0.93]$, $p < 0.001$, for the 4-year-olds; and 0.93, 95% CI $[0.83, 0.98]$, $p < 0.001$, for the 5-year-olds. These results indicate that most children understood the game and could reliably voice out or point toward their choice, providing a solid foundation for proceeding with the analyses of the test trials.

Moving on to the main analyses, we first assessed how well the three age groups performed overall in the 16 test trials. As can be seen in Fig. 3, in line with the hypothesis of a developmental trend of probabilistic reasoning, older participants tended to provide more correct responses than younger ones. More precisely, the average number of correct responses increased steadily with age: from $M = 10.17$, 95% CI $[9.19, 11.1]$ (corresponding to 63.60% accuracy) in the 3-year-olds, to $M = 11.27$, 95% CI $[10.29, 12.2]$ (70.40% accuracy) in the 4-year-olds, and to $M = 12.87$, 95% CI $[11.89, 13.8]$ (80.40% accuracy) in the 5-year-olds. Additionally, the graph's density and dot plots show that the distribution of participants reaching high accuracy is right-skewed in the 5-year-old group whereas it is not in the younger age groups.

We conducted a global factorial ANOVA using the afex package (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2022), which tested whether performance accuracy differed in the different age groups. This analysis yielded a significant effect of age, $F(2, 87) = 7.57$, $p < .001$, $\eta_p^2 = .15$. Following up on this main effect, we conducted a polynomial contrast analysis to test if performance followed the predicted developmental trend. Corroborating this prediction, and consistent with the descriptive pattern shown in Fig. 3, this analysis yielded a significant linear trend in participants' accuracy, $t(87) = 3.87$, $p < .001$.

Given the observed age differences in overall accuracy, we also wanted to determine what group performed above chance level (i.e., 50% [8 out 16] correct choices) and whether their performance could

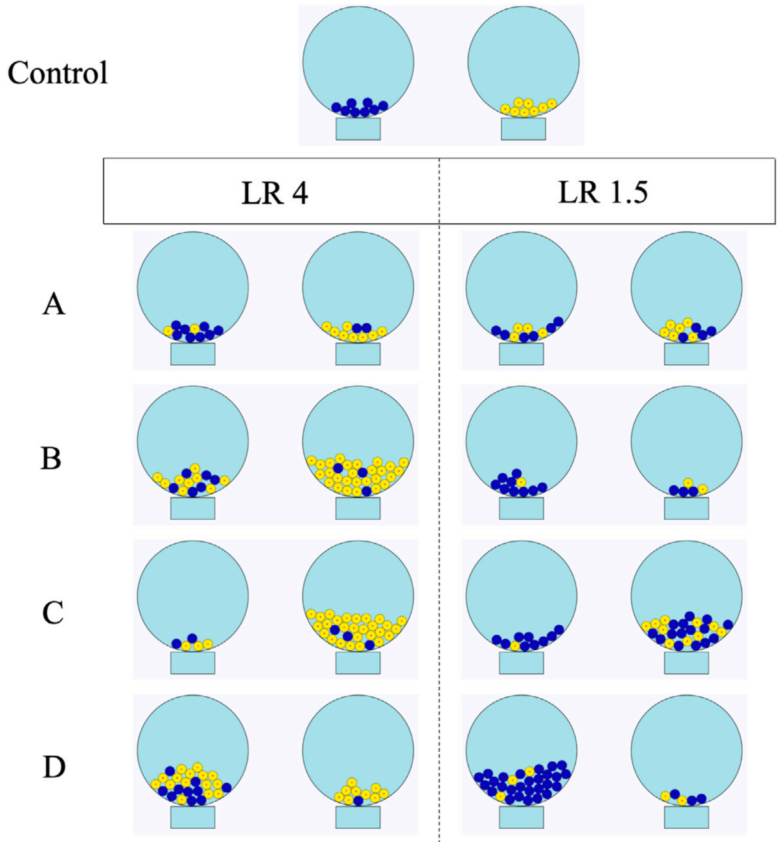

Fig. 2. Examples of the urns presented in the experiment, ordered by trial type and LR.
Note. In the examples above, the target color is blue and, therefore, the more favorable urn is always on the left (although, in the experiment, the target color was counterbalanced across children and the side of the most favorable urn was counterbalanced across trials). Control trials have a deterministic outcome and served to ensure that children understood and engaged with the task. In type A trials, all strategies led to the correct answer, except for the 'Random' strategy, which was expected to lead to the correct answer only in half of the cases. In type B trials, all strategies led to the correct answer, except for the 'Random' and the 'Good urn' strategies, which were expected to lead to the correct answer only in half of the cases. In type C trials, the 'Ratio' (i.e., the proper probabilistic inference) and the 'Less bad' strategies led to the correct answer, the 'More good' strategy led to the wrong answer, while the 'Random' and the 'Good urn' strategies were expected to lead to the correct answer only in half of the cases. Finally, in type D trials, the 'Ratio' and the 'More good' strategies led to the correct answer, the 'Less bad' strategy led to the wrong answer, while the 'Random' and the 'Good urn' strategies were expected to lead to the correct answer only in half of the cases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

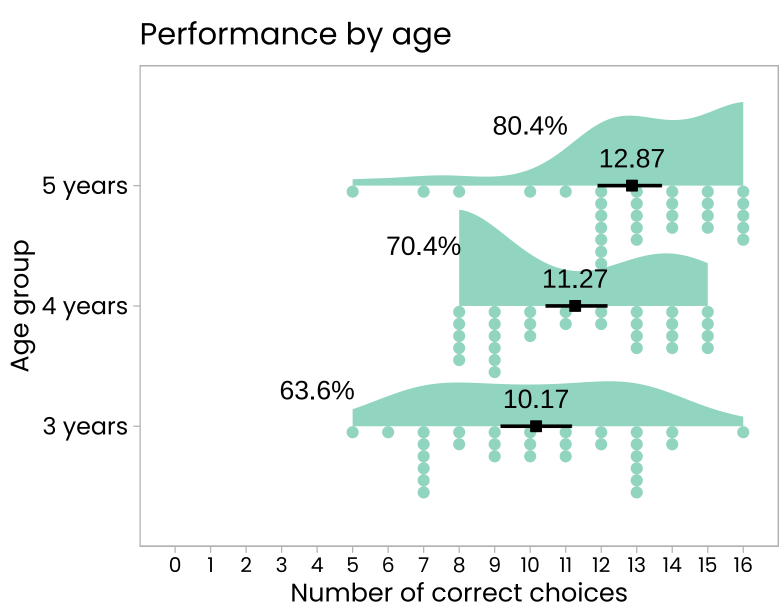

Fig. 3. Participants' performance (number of correct predictions) in the different test trials of the experiment.
Note. Overall performance across the 16 test trials for the three age groups. Black squares show means and error bars of their 95% CI. Stacked dots represent individual participants. Density plots show the distribution of participants' choices.

be better explained by heuristic strategy or proper probabilistic reasoning.

To address the first question, we conducted separate two-sided one-sample proportion tests (with continuity correction) using the prop.test function from R's stats package. The result was that all three age groups performed better than chance (3-year-olds: 63.6%, $\chi^2[1] = 34.67$, $p < .001$; 4-year-olds: 70.4%, $\chi^2[1] = 79.22$, $p < .001$; 5-year-olds: 80.4%, $\chi^2[1] = 176.42$, $p < .001$). Crucially, this indicates that none of the three age groups completely relied on the 'Random' strategy. However, it does not necessarily imply that children relied on normative probabilistic reasoning. Indeed, as discussed above, relying on some of the heuristic strategies considered could lead to decent performance.

The two heuristic reasoning strategies with the highest prediction accuracy in our test trials are the 'More good' and the 'Less bad' strategy (cf. Table 2). They both predict the correct urns in 75% of the test trials. The 3- as well as the 4-year-olds had a prediction accuracy that was lower than that (63.6% and 70.42%, respectively) but the performance of the 5-year-olds was above that (80.40%). A two-sided proportion test testing the 5-year-olds average prediction accuracy against a level of 75% correct choices showed that their performance accuracy of 80.42% was significantly higher than that, $\chi^2(1) = 7.23$, $p = .0036$. These analyses suggest that while the overall performance level of the 3- and 4-year-olds may plausibly be explained by simpler heuristic reasoning strategies, the high overall performance accuracy of the 5-year-olds may reflect a more advanced quantification of probabilities, primarily based on proportions.

We next analyzed if and to which extent performance was influenced by the other factors we considered in the study. To test if LR and trial type make a difference in addition to participants' age, we fitted and compared different logistic generalized linear mixed effects models (GLMM) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was whether the correct response was given in a given test trial. We fitted a null model with individual IDs implemented as a random effect and then, step-wise, added age, LR, and trial type as predictors, their two-way interactions, and the three-way interaction between age, LR, and trial type. We also fitted a model that included urn size. The different models were compared using R's ANOVA function.

The analysis yielded a significant interaction effect between LR and trial type, $\chi^2(2) = 11.41$, $p = .0097$. Following up on this interaction effect with post-hoc comparisons, we found that type A and C trials were harder than type B and D trials when LR was smaller (LR = 1.5) but not when LR was higher (LR = 4). (The test statistics for these comparisons can be found in the analysis script provided at the repository site.) The interaction pattern we found is shown in the interaction plot in Fig. 4.

As can be seen in the plot, children's performance in type C trials was generally low (i.e., 66% when LR was 1.5 and 67% when LR was 4) and did not seem to be affected by the LR. Part of what seemed to have driven the interaction effect was the drop in performance in type A trials when LR was 1.5. Children's low performance in type C trials was predicted and could be explained by the fact that many children relied on the 'More good' strategy, which in type C trials would lead to the wrong answer. One possible explanation for the drop in performance in type A trials with LR 1.5, which was not predicted, is that these two trials (i.e., trials 9 and 10, see Table 1) were among the few trials in which the ratio of target objects was exactly 2:3 (i.e, in trial 9, there were 6 target balls in one urn and 4 in the other urn; in trial 10, there were 8 target balls in one urn and 12 in the other urn). In all other trials, this ratio was higher (except in the two trials of Type C with LR 4, cf. Table 1). This means that in the two trials of type A with LR 1.5, the two urns were harder to differentiate regarding their absolute number of target items. This small ratio would, however, only hinder children who rely on the 'More good' strategy because this is the only strategy that relies on comparing the absolute number of target objects to pick an urn. As mentioned above, we chose a ratio of 2:3 as the minimum ratio between the numbers of elements that needed to be compared according to the different strategies because previous findings suggested that much younger children could discriminate such small ratios (Xu & Arriaga, 2007). Thus, we wanted to ensure that children could compare all the quantities necessary to use the different strategies we wanted to monitor. It is possible, however, that the presence of non-target objects in the urns made comparing a 2:3 ratio of target objects harder, explaining children's poorer performance in type A trials with LR 1.5.

Finally, urn size did not have a significant effect on performance accuracy.

3.3.2. Exploratory group-level analyses

Another aspect we wanted to explore was the possibility of a learning effect, meaning an increase in performance throughout the experiment. A learning effect would imply that participants' probability of giving the correct answer in a given test trial increased with the number of test trials they had already encountered. To investigate this, we first fitted a GLMM that included the 'trial order' factor (i.e., the order in which participants were presented with a specific test trial) in addition to the previous predictors. When we compared this model to the previous ones, we found that trial order significantly affected performance, $\chi^2(1) = 12.50$, $p < .001$. A graphical summary showing how exactly the proportion of correct responses in the three different age groups changed during the course of the experiment is provided in Fig. 5. As can be noted, consistent with a learning effect, the probability of giving the correct answer tended to increase as participants completed more trials. However, this learning effect was most pronounced in the 5-year-olds, less so in the 4-year-olds, and nearly absent in the 3-year-olds. Accordingly, the correlations between trial order and proportion of correct responses were significantly different from zero only in the 5- and 4-year-olds but not in the 3-year-olds (see test statistics in Fig. 5).[4]

Did learning occur because participants began associating the correct quantity (i.e., proportions) with the feedback they received after each trial? To see if trial feedback drove the learning effect, we fitted a generalized linear model for each age group that predicted participants' correct responses from those in the previous trial and the outcome of the previous trial, and we implemented participant IDs as random effect. Unlike what would be expected if the effect resulted from simple trial-and-error learning, neither the previous choice, the previous outcome, nor their interaction were significant predictors of participants' performance (see Table 3).

In a final exploratory analysis, we also examined whether gender might influence performance. A GLMM that included gender along with all previous factors did not identify it as a significant factor, which aligns with previous findings about gender effect in math cognition (Hutchison, Lyons, & Ansari, 2019).

In sum, these group-level results align with the predicted developmental trend in probabilistic reasoning: Older participants performed better than younger ones, and their performance also improved more during the experiment than that of younger participants. The finding that trial type influenced participants' performance suggests that a substantial proportion of participants, particularly younger ones, relied on quantity heuristics. In particular, the lower performance in type C trials further indicates that one of the dominant heuristic strategies used by these participants was the 'More good' strategy.

[4] It should be noted, however, that the correlation coefficients of the 5- and 3-year-olds did not differ significantly from each other when compared in two-sided tests of difference between two independent correlations. Furthermore, when comparing a GLMM that also included the interaction between trial order and age to a model that only included their main effects, the interaction did not result significant, both when the trial order was entered into the analyses as a factor with 16 levels or as a factor with two levels (i.e., the first and last halves of the trials). The tests are reported in the analysis script.

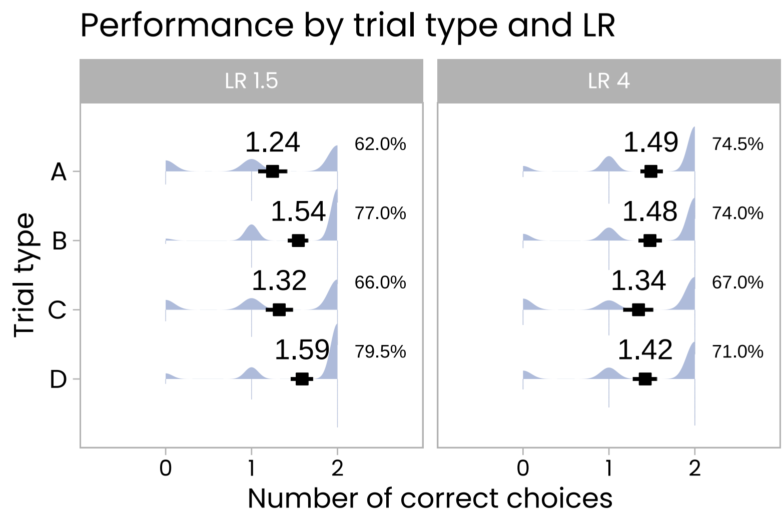

Fig. 4. Participants' performance (number of correct responses) in the different trial types of the experiment.
Note. The two panels represent the two likelihood ratio conditions. Black squares, numbers, and percentages represent mean performance accuracy. Error bars represent 95% CIs of the means. Stacked dots represent individual participants. Density plots show the distribution of participants' correct responses for each trial type.

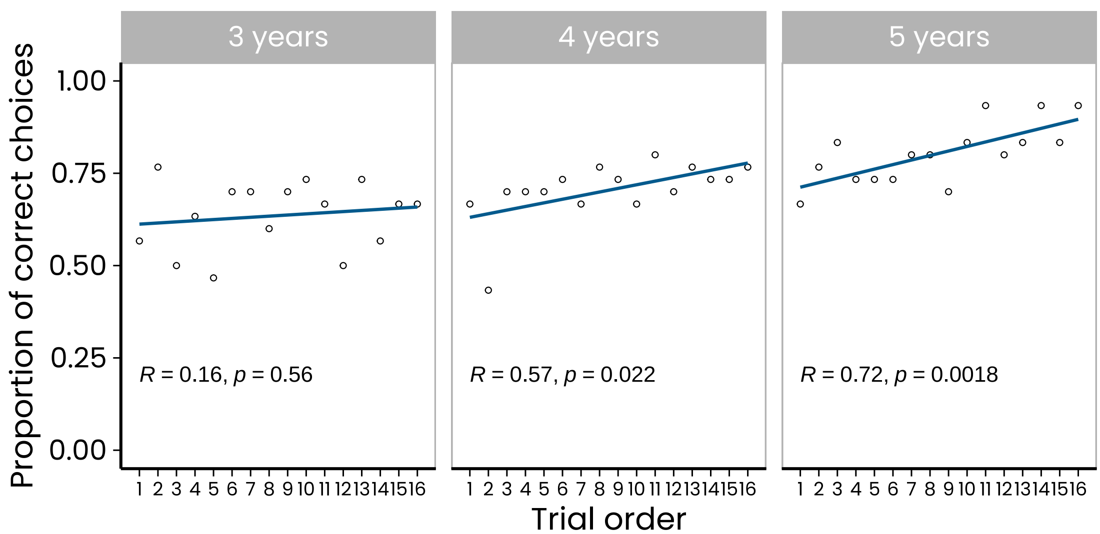

Fig. 5. Proportion of correct responses in the three age groups as a function of the number of test trials already encountered.
Note. Blue lines represent regression lines and the equations show correlation coefficients together with their corresponding p values.

Table 3
Results of the logistic regression model predicting participants' correct responses from their responses in the previous trial and the outcome obtained in the previous trial.

	Estimate	Std. error	z value	p-value	
Intercept	1.134	0.082	13.805	0.000	***
Previous feedback negative	0.115	0.159	0.725	0.468	
Previous response wrong	−0.224	0.222	−1.007	0.314	
Previous feedback negative: Previous response wrong	−0.322	0.288	−1.118	0.264	

3.3.3. Individual-level analyses

Unlike most previous studies that mainly focused on overall group-level performance, we also aim to explore the different reasoning strategies individual participants may have used (but see Falk et al., 2012, for another study in which individual strategies were considered). Looking at individual-level strategies is informative because even though the groups of 3- and 4-year-olds, as a whole, seemed to primarily rely on heuristic strategies, there might be a non-negligible number of small children that were able to engage in proper probabilistic reasoning. Also, we wanted to see in which proportions the different strategies were used by the different groups. For this reason, we examined which strategy was most consistent with each participant's responses across the 16 test trials. In other words, we identified the dominant strategy (or, possibly, combination of strategies) that each participant relied on the most during the experiment. We then used this information to refer back to the group level by calculating the proportions of the dominant strategies within the three different age groups.

To identify a participant's dominant strategy across the 16 test trials, we calculated the match between their responses in each trial and the expected choice pattern for each of the reasoning strategies considered. We then assigned the participant the strategy with the highest overall match. If two or more strategies tied (e.g., both accounted for 10 judgments), we assigned all of them to the participant but weighing

Table 4
Dominant strategy percentages within each age group, ordered from highest percentage to lowest.

Dominant strategy	3 years	4 years	5 years
Ratio	44.2	60.6	67.8
More good	40.8	20.6	16.1
Less bad	10.8	12.2	12.8
Random	3.3	6.7	3.3
Good urn	0.8	0	0

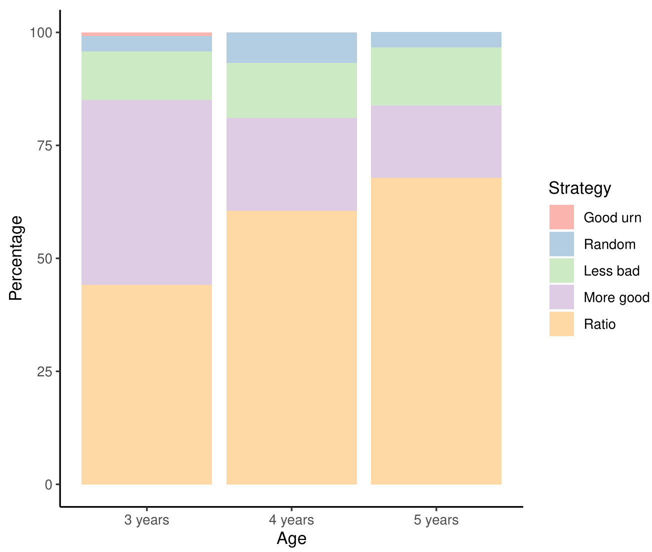

Fig. 6. Percentages of dominant strategies within each age group.

them according to the number of tied strategies. Finally, we calculated the proportion of dominant strategies for each age group (see the analysis script for further details on how the dominant strategies were identified).

The results are presented in Table 4 and graphically illustrated in Fig. 6. As can be seen there, we found that most of the participants' choice behaviors was best described by the normative 'Ratio' strategy in all age groups (although in the 3-year-olds, there was not much of a difference between their use of the 'Ratio' and the 'More good' strategies). This suggests that the ability to represent and use proportions may already be present from a young age. Also, in line with a developmental trend of probabilistic reasoning, we found that the prevalence of the 'Ratio' strategy steadily increased across age groups: it was primarily used by 44.2% of 3-year-olds, by 60.6% of 4-year-olds, and by 67.8% of 5-year-olds. The second most frequent strategy was the 'More good' strategy in all age groups. This aligns with previous research, as well as with our finding that performance tended to be worse in type C trials, where this strategy leads to incorrect answers. Also in line with a developmental trend of normative probabilistic reasoning is the finding that the prevalence of this heuristic reasoning strategy steadily decreased with age: while it was the dominant strategy for 40.8% of the 3-year-olds, this number dropped to 20.6% of the 4-year-olds and to 16.1% of the 5-year-olds. The third most frequent dominant strategy was the 'Less bad' strategy, whose frequency did not vary much across ages (10.8% of 3-year-olds, 12.2% of 4-year-olds, and 12.8% of 5-year-olds seemed to primarily rely on it). The fourth most frequent dominant strategy was the 'Random' strategy, which was generally used less frequently (3.3% of 3-year-olds, 6.7% of 4-year-olds, and 3.3% of 5-year-olds seemed to primarily rely on it, which corresponds to one 3-year-old, two four-year-olds, and one five-year-old). Finally, the 'Good urn' strategy was the least frequent dominant strategy, and never exceed 1% overall (0.1% of 3-year-olds,[5] 0% of 4-year-olds, and 5-year-olds seemed to primarily rely on it).

The results of individual-level analyses complement those from group-level analyses, indicating that proper probabilistic reasoning gradually develops with age. While it can be present in some 3-year-olds, it becomes more prevalent by age 5. Notably, this finding emerged only through our individual-level analyses. Regarding the heuristic strategies, our results confirm that the most frequently used heuristic is the 'More good' strategy and that its use diminishes with age in favor of the 'Ratio' strategy, that is, a proper probabilistic inference.

4. General discussion

This paper investigated the ontological development of probabilistic reasoning using a refined version of the widespread two-urn paradigm. In 16 test trials, participants were presented with different combinations of urns containing varying proportions of blue and yellow balls. To maximize their chances of winning a sticker, they had to predict which urn was more likely to yield a target ball after several random mixtures of the urns' content. We took great care to disentangle choices driven by normative probabilistic reasoning from those based on less accurate heuristic reasoning strategies. In our task, while choosing the urn with the higher proportion of target objects (i.e., using the 'Ratio' strategy) would lead to correct responses in all trials, relying on quantity heuristics (i.e., the number of target or non-target objects) would result in correct answer only in certain trials. Specifically, the quantity heuristics we controlled in our study were the 'More good' strategy (i.e., choosing the urn with more target objects), the 'Less bad' strategy (i.e., choosing the urn with fewer non-target objects), and the 'Good urn' strategy (i.e., choosing a good urn – i.e., an urn with more target than non-target objects – over a bad urn – i.e., an urn with more non-target objects than target-objects). In addition, we also controlled whether participants chose randomly between urns (the 'Random' strategy).

Our findings revealed a developmental trend in children's probabilistic reasoning skills: the older the children, the more likely they were to rely on comparing proportions of target objects; the younger they were, the more likely they were to compare absolute numbers of objects. However, an additional individual-level analysis, which categorized participants according to the strategy (or combination of strategies) they most likely used, revealed that even some (i.e., about half) of the youngest participants relied on comparing proportions. The ability to quantify probabilities based on proportions and use them to guide decisions is, therefore, already within reach of 3-year-olds. However, at the group level, a significant number of children rely on comparing proportions to compute probabilities only by age 5.

Why did most of the younger preschoolers fail to provide the correct answer while an appreciable number of 5-year-olds performed well? We can, with some confidence, rule out the possibility that younger children failed to identify the sample space, as most of them correctly selected the urn in the control trials, in which the outcomes were deterministic, and did, therefore, not appear to expect balls to be drawn from outside the urns. Furthermore, all children seemed to understand that urns containing different combinations of target and non-target balls were associated with different probabilities of drawing a target ball, as evidenced by their non-random choices in the test trials, in which the outcomes were non-deterministic. The ability to compare proportions of objects should not be the only main concern, as previous studies (see, for example, McCrink & Wynn, 2007) indicate that even younger children than those in our study can perform this task at least when the two LRs differ by a factor of 4 (i.e., in half

[5] This corresponds to 0.03 children, which is not a possible amount of children. It means that this strategy always tied with other strategies and was never assigned as a unique dominant strategy.

of our trials). This does not imply that explicit comparisons between proportions are always carried out correctly. However, it is plausible that (at least part of) the explanation of our results lies in the process of translating proportions into relevant probability values. In this regard, it is important to emphasize that estimating the probability of obtaining a target ball from a specific urn depends on its proportion of target balls under the assumption that the draw is random. This assumption, in turn, requires understanding the consequences of the mixing of the balls during the animation. More specifically, to provide the correct response, the children in our study needed to understand that they could not deterministically predict the movements of the balls. The younger ones may have struggled to grasp that the distribution of balls within the urns could be interpreted as probability distributions that could be used to estimate the likelihood of specific events (i.e., the outcome of a particular draw). Faced with uncertainty and unable to discern its laws or rules, they might have resorted to simpler strategies, such as quantity-based heuristics. Piaget and Inhelder (1975) already provided a similar explanation during their investigation of children's concept of chance. Interestingly, they found that the understanding of the consequences of random mixtures develops alongside the ability to correctly infer the outcomes of random draws. Although they attributed this ability to older children, likely due to the higher difficulty of their tasks (see Introduction), the link they established between understanding random mixtures and the outcomes of random draws could still hold.

Consistent with previous studies (Falk et al., 2012; Goldberg, 1966; Piaget & Inhelder, 1951), our results confirmed that the most commonly used quantity heuristic among young children is the 'More good' strategy. This quantity heuristic, also known as *Denominator neglect*, appears to persist into adulthood, as individuals also tend to focus on numerators when comparing risks in health-related situations (see, for example, Garcia-Retamero, Galesic, & Dhami, 2012; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Reyna & Brainerd, 2008). A potential explanation for this behavior could be attributed to an intuitive rule (also referred to as 'Attribute substitution' (Kahneman & Frederick, 2002), which is commonly used to solve complex tasks in physics and mathematics (Falk et al., 2012; Stavy & Tirosh, 1996). This involves answering difficult questions (sometimes without even realizing it) by rephrasing them into a simpler form that appears equivalent but is not truly so. In particular, when asked to compare objects based on a less obvious variable B, children (and adults in some situations) may rely on another more salient variable A, applying the rule 'The more of A, the more of B'. For example, if children younger than 6 are asked to compare the quantity of water contained in two containers, they tend to use the rule "The higher the level of water, the more water". This rule, however, only works if the two containers have equal dimensions. In our experiment, the absolute number of target items was a salient quantity that many children may have relied on when faced with the difficulty of the task.

Our data also show an improvement in the performance of 4- and 5-year-olds throughout the experiment, indicating that they learned something. Although this exploratory finding should be treated cautiously, especially since the exact nature of the children's learning is uncertain, it remains an intriguing observation. One possible explanation for the observed learning effect is that children needed time either to understand that this was a probability game (which required rational answers rather than mere guessing) or to become more proficient at distinguishing proportions of target objects. Even more intriguing, given the general difficulty of improving probabilistic reasoning also in adults, is the possibility that the feedback helped children better grasp the concept of probability and identify the most effective rule for maximizing their rewards. In this regard, it is important to note that the feedback was genuine (i.e., was probabilistic and not deterministic) and that our data show that previous outcomes did not affect children's subsequent answers (even when their correct responses were followed by negative outcomes). This suggests that (at least most of older) children not only understood that probabilities were based on proportions but also comprehended the probabilistic nature of the reward (i.e., that a correct choice did not guarantee a favorable outcome). Such a deeper understanding of probabilities has previously been reported only in older children around 10–11 years old (Falk et al., 2012).

4.1. Comparison to previous findings

The methods we used in this study were similar to those of Gualtieri and Denison (2019), who also tested preschoolers with virtual containers. However, our findings diverge. Their results show a better performance of 3- and 4-year-olds, regardless of trial type, and did not indicate any learning effect. One reason for the better performance of 3- and 4-year-olds in their study could be that a good performance could also be reached without assessing proportions. As mentioned in our introduction, most trials could have been solved using a 'Good urn' strategy. The trials that could not be solved with that strategy could have been solved using modal concepts rather than probabilities, i.e., children could have chosen the necessary over the possible outcome, or the possible over the impossible one (see our Introduction).

One could argue that, given that children seem not to have used the 'Good urn' strategy in our study, there is no reason they did in Gualtieri and Denison (2019)'s experiments. It is possible that the children in our study did not rely on the 'Good urn' strategy because it was mostly inconclusive. In fact, in most of our trials, both urns were either two 'Good' urns or two 'Bad' urns. We surmised that children who relied on the 'Good urn' strategy would choose at chance in such a case. However, as said above, random choice was not prevalent in our data. Instead of choosing randomly, small children may have adopted a more effective strategy (i.e., the 'More good' strategy). In Gualtieri and Denison (2019), the selection patterns of participants following the 'Good urn' strategy would have aligned with the 'Ratio' strategy (except for the trials that could be solved with modal concepts instead of probabilities). In light of our own findings, this could mean that their study overestimated the number of participants who solved the task by engaging in proper probabilistic reasoning. If most participants in Gualtieri and Denison (2019) relied on the 'Good urn' strategy, it could also explain why their performance did not improve during the game, as differentiating good from bad urns is a relatively easy task that might not require much learning. Alternatively, an explanation for why participants' performance did not improve in Gualtieri and Denison (2019) could be that they underwent only 8 trials. If we rerun our generalized linear mixed effects model considering only the first 8 trials, our learning effect disappears (see our analysis script).

Our findings also differ from those of Falk et al. (2012). While we found that probabilistic reasoning develops between 3 and 5, Falk et al. (2012)'s findings suggest that it develops around 7 and that children tend to rely on the 'More good' strategy until then. As mentioned above, their results could have partly been driven by the small LRs they used in their study. It is not implausible to assume that reasoners turn to use heuristic strategies when proportions become too hard to compare. In our study, the two LRs employed did not affect performance. One reason for this might be that a significant number of small children relied on heuristic strategies. Another possibility is that the LR of 1.5 was not low enough to affect the performance of children who wanted to use proportions. Falk et al. (2012) used LRs lower than 1.5 in more than 70% of their trials. In fact, in some of their trials, the LR was so small that solving the task would have necessitated explicit counting (see, for example, Trial 21 of Table 3 in Falk et al. (2012)).

Our findings also do not align fully with those of Girotto et al. (2016) and Placì et al. (2020), which, taken together, concluded that 3- and 4-year-olds cannot compute probabilities based on proportions. In particular, their findings suggested that 3-year-olds were not relying on any quantitative information, as they performed at chance level in almost all probability conditions, even when these could have been easily solved by heuristic strategies or modal concepts. One possibility for 3-year-olds' failure to use the provided quantitative information to draw inferences about the objects removed from the urns could be due to the epistemic nature of the uncertainty they had to deal with (Robinson et al., 2006) or to the difficulty in quantifying mixtures of 3-D objects (see Introduction). It is also possible, however, that participants of these studies did not consider the game as a probabilistic game at all. Indeed, the methods used by Girotto et al. (2016) and Placì et al. (2020) only simulated random draws: The experimenters pretended to pick an item from the containers randomly. In reality, they always selected the object type whose proportion was highest. The children could have believed that the experimenter was playing tricks or picking their own favorite color, which would have turned the game into a guessing game. To control if this was the case, Girotto et al. (2016) did a third experiment with only 3-year-olds, in which the objects were inserted into two plastic urns that ended in an opaque neck. This time, the experimenter did not perform the draw; they performed the mixing of the objects with a little stick, however. It is possible that this was not sufficient to make children understand that *any* object could fall out of the urn. Another reason that could explain the divergence between our findings and those of Girotto et al. (2016) is that the poor performance of 3-year-olds in their studies hides two sub-groups, one group of children who can use probabilities and another that relies on quantity heuristics. The authors ruled that possibility out, given that only a few children in their study made no errors and most reached an intermediate level of performance. However, it is possible that some of the children who made some mistakes would still have predominantly used a 'Ratio' strategy if they had enough trials to figure out which rule worked best in this game. Whatever the actual reason that explains 3-year-olds' failure in Girotto et al. (2016) and Placì et al. (2020), we deem that the methods used in the current experiment allowed a more comprehensive evaluation of preschoolers' probabilistic reasoning abilities or, at least, their ability to use specific quantity heuristics.

Although our findings indicate that (some) 3-year-olds can compute probabilities based on proportions, our results are hard to reconcile also with those of Denison and Xu (2014), which indicate that 12-month-olds are capable of probabilistic reasoning. In our study, many 3- and 4-year-olds still seemed to rely on heuristic strategies. In contrast, in Denison and Xu (2014), most infants chose the object drawn from the urn with a higher proportion of target objects, even when the absolute number of target objects was higher in the other urn. Probabilistic reasoning could follow a U-shaped trend in development (Pauls, Macha, & Petermann, 2013). However, this explanation seems not only convoluted, but also unsupported, as a replication of the study by Placì et al. (2020) yielded different results. However, in this replication, infants' performance was at chance level in the deterministic condition in which probabilities need not be used, which makes the results harder to interpret as it is possible that participants did not understand the procedure. Further attempts to replicate Denison and Xu (2014)'s study are needed to reach a clearer conclusion about the probabilistic reasoning ability of infants.

A potential limitation of our study is that, unlike previous research on probabilistic reasoning in preschoolers, it was conducted through an online meeting platform due to the restrictions imposed by the COVID-19 pandemic. Previous research (Lapidow, Tandon, Goddu, & Walker, 2021) found that the performance of younger preschoolers (i.e., 3- and 4-year-olds) in online reasoning tasks follows the same trend as in in-person procedures, though it may be weaker. In principle, it is then possible that the results obtained with the youngest children in our study were, at least in part, affected by the online setting. If this were the case, it could explain why they did not perform as well as their peers in Gualtieri and Denison (2019). However, this hypothesis is challenged by the strong performance of the 3- and 4-year-olds in the control trials of our study, with 92% and 85% correct answers, respectively, which closely align with the figure (88%) reported by Gualtieri and Denison (2019) in a similar control trial (see Problem D in Experiment 1, p. 382). These data support the hypothesis that our online study did not impose more task demands on younger children than previous in-person studies.

4.2. Directions for future research

Our findings suggest that the ability to compute probabilities based on proportions is present in a minority of 3- and 4-year-olds but fully develops after 5 years. Future studies could investigate the abilities of younger children (e.g., two-year-olds) with the same method we used. In our experiment, participants could point toward the urns to indicate their preference, which makes the task accessible to younger children. Experiments with younger children could shed light on the heuristic strategies that are more prevalent at a younger age and provide further insights into the age at which probabilities begin to be computed from proportions.

Future studies could also explore more complex probabilistic reasoning tasks in preschoolers, such as updating judgments. Previous studies suggest that this ability develops around the age of 5 (Girotto & Gonzalez, 2008), which matches the age at which the ability to compute probabilities based on proportions seems to stabilize in most children. However, it would be interesting to test whether Bayesian updating is already present in children who can rely on proportions to compute probabilities earlier in their development.

Follow-up studies could look into the mechanisms underlying probabilistic reasoning. One of the components of probability is *randomness* (Bryant & Nunes, 2012; Nikiforidou & Pange, 2010; Supply, Wijns, Van Dooren, & Onghena, 2023). Thus, understanding randomness should be a prerequisite for the understanding and the correct quantification of probabilities (Bryant & Nunes, 2012). Studies examining children's understanding of randomness suggest, however, that this notion is hard to grasp, even more so in younger children (Nikiforidou & Pange, 2010; Supply et al., 2023), and that young children typically expect some kind of order to appear out of random events. How do we reconcile these findings with those that indicate that preschoolers, and at least 5-year-olds, can reason about probability?

Additional research could also focus on the link between the development of modal concepts and probabilistic reasoning (see, for example, Alderete & Xu, 2023). Some studies suggest that modal concepts are only minimally represented in 3-year-olds (Leahy, 2023; Leahy & Carey, 2020). In contrast, others suggest that children, even during infancy, compute probabilities based on an enumeration of the possibilities in which an event can unfold (Téglás et al., 2015). Better understanding the relationship between reasoning about possibilities and probabilities could help us better comprehend the development of decision-making under uncertainty.

4.3. Conclusion

Previous research yielded contrasting findings regarding preschoolers' ability to compute probabilities based on proportions. We set out to identify the reasons for these discrepancies and refined a well-known experimental paradigm to collect more reliable data on the probabilistic reasoning skills of 3- to 5-year-olds. We found that, while the ability to compute probabilities based on proportions is already present in some 3-year-olds, it fully develops at around 5. Most children before that age rely on various heuristic strategies, the most prominent of which is comparing the absolute number of target items rather than proportion. Our findings also highlight the importance of analyzing individual strategies and group-level performances to get a comprehensive picture of the development of children's reasoning abilities.

CRediT authorship contribution statement

Sarah Placì: Conceptualization, Writing – original draft, Methodology, Formal analysis, Project administration, Funding acquisition, Writing – review & editing, Visualization. **Stefania Pighin:** Methodology, Supervision, Writing – review & editing. **Tommaso Mastropasqua:** Writing – review & editing, Software. **Katya Tentori:** Methodology, Writing – review & editing, Formal analysis, Supervision.

Declaration of competing interest

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

Acknowledgments

We thank all parents and children who participated in this study. We thank Simon Stephan for his helpful comments on the manuscript.

This research has been funded with an Early Postdoc.Mobility fellowship received from the Swiss National Science Foundation with grant No P2SKP1_191311, as well as with a Postdoc.Mobility fellowship received from the Swiss National Science Foundation with grant No P500PS_206568. This research was also funded by the German Research Foundation (DFG) in the form of a Reinhart Koselleck project (WA 621/25-1).

Data availability

All experimental materials and data are provided in a GitHub repository whose content can be accessed via a GitHub page at https://sarah-placi.github.io/prob_reasoning_preschoolers/index.html.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear Mixed-Effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Bryant, P., & Nunes, T. (2012). *Children's understanding of probability: A literature review (full report)*: *A report to the nuffield foundation*.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5-and 6-month-olds. *Developmental Psychology*, *49*(2), 243–249.

Denison, S., & Xu, F. (2010). Twelve-to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, *13*(5), 798–803.

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

Denison, S., & Xu, F. (2019). Infant statisticians: The origins of reasoning under uncertainty. *Perspectives on Psychological Science*, *14*(4), 499–509.

Falk, R., Yudilevich-Assouline, P., & Elstein, A. (2012). Children's concept of probability as inferred from their binary choices—revisited. *Educational Studies in Mathematics*, *81*, 207–233.

Garcia-Retamero, R., Galesic, M., & Dhami, M. K. (2012). Reducing denominator neglect. *Transparent Communication of Health Risks: Overcoming Cultural Differences*, 145–164.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

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

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition*, *106*(1), 325–344.

Goldberg, S. (1966). Probability judgments by preschool children: Task conditions and performance. *Child Development*, 157–167.

Gualtieri, S., & Denison, S. (2019). A comprehensive examination of preschoolers' probabilistic reasoning abilities. In *CogSci* (pp. 380–386).

Hutchison, J. E., Lyons, I. M., & Ansari, D. (2019). More similar than different: Gender differences in children's basic numerical skills are the exception not the rule. *Child Development*, *90*(1), e66–e79.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, *49*(49–81), 74.

Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A tale of three platforms: Investigating preschoolers' second-order inferences using in-person, zoom, and lookit methodologies. *Frontiers in Psychology*, *12*, Article 731404.

Laplace, P.-S. (1814). Philosophical essay on probabilities. Translated from the fifth french edition of 1825. In *Sources in the history of mathematics and physical sciences (book 13)*.

Leahy, B. P. (2023). Don't you see the possibilities? Young preschoolers may lack possibility concepts. *Developmental Science*, *26*(6), Article e13400.

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

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, *18*(8), 740–745.

Nikiforidou, Z., & Pange, J. (2010). The notions of chance and probabilities in preschoolers. *Early Childhood Education Journal*, *38*, 305–311.

Pauls, F., Macha, T., & Petermann, F. (2013). U-shaped development: An old but unsolved problem. *Frontiers in Psychology*, *4*, 301.

Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant*. Presses Universitaires de France.

Piaget, J., & Inhelder, B. (1975). *Psychology revivals*, *The origin of the idea of chance in children*. Routledge and Kegan Paul, London.

Placì, S., Fischer, J., & Rakoczy, H. (2020). Do infants and preschoolers quantify probabilities based on proportions? *Royal Society Open Science*, *7*(9), Article 191751.

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107.

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child Development*, *77*(6), 1642–1655.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). Afex: Analysis of factorial experiments. URL https://CRAN.R-project.org/package=afex, R package version 1.1-1.

Stahl, A. E., & Feigenson, L. (2024). Young children distinguish the impossible from the merely improbable. *Proceedings of the National Academy of Sciences*, *121*(46), Article e2411297121.

Stavy, R., & Tirosh, D. (1996). Intuitive rules in science and mathematics: the case of 'more of A–more of B'. *International Journal of Science Education*, *18*(6), 653–667.

Supply, A.-S., Wijns, N., Van Dooren, W., & Onghena, P. (2023). It is probably a pattern: does spontaneous focusing on regularities in preschool predict reasoning about randomness four years later? *Educational Studies in Mathematics*, *112*(1), 3–24.

Téglás, E., & Bonatti, L. L. (2016). Infants anticipate probabilistic but not deterministic outcomes. *Cognition*, *157*, 227–236.

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, *104*(48), 19156–19159.

Téglás, E., Ibanez-Lillo, A., Costa, A., & Bonatti, L. L. (2015). Numerical representations and intuitions of probabilities at 12 months. *Developmental Science*, *18*(2), 183–193.

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

Xu, F., & Arriaga, R. I. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology*, *25*(1), 103–108.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97–104.

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

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(1), 28–32.

Yost, P. A., Siegel, A. E., & Andrews, J. M. (1962). Nonverbal probability judgments by young children. *Child Development*, 769–780.