A comprehensive examination of preschoolers' probabilistic reasoning abilities

Samantha Gualtieri (sgualtieri@uwaterloo.ca) & Stephanie Denison (stephanie.denison@uwaterloo.ca)
Department of Psychology, University of Waterloo
Waterloo, Ontario, Canada

Abstract

Historically, research on preschool-aged children's probabilistic reasoning abilities has yielded mixed results. Although some findings have suggested that young children can successfully evaluate probabilities, others have suggested that they may use strategies that only approximate true probabilistic inference and therefore sometimes make errors (e.g., Girotto et al., 2016; Piaget & Inhelder, 1975). To explore the factors that affect young children's probabilistic reasoning, we developed a battery of problems that contained features that affect the ease with which a problem is evaluated, and the types of alternative strategies that can be applied to solve them. The current experiments (total $N = 124$) assessed 3- and 4-year-old children's probabilistic reasoning using an experimental paradigm tailored to this age group. Results from both experiments suggest that young children are able to engage in true probabilistic inference, as they performed well-above chance on each problem. Nuances in children's performance are discussed, along with possibilities for future research.

Keywords: probabilistic reasoning; cognitive development; decision making

Introduction

Our ability to make inferences under uncertainty is critical to learning and decision-making, mimicking the contexts in which every day reasoning tends to occur. That is, we are typically in situations where we only have access to probabilistic information. Although sensitivity to base-rates is evident very early in development, questions remain regarding young children's strategies for using base-rates in their inferences, which can be diagnosed by asking children to make more complex proportional comparisons.

Non-human primates, infants, toddlers, and preschoolers correctly infer that an object of the majority type is most likely to be randomly sampled from simple probabilistic distributions (Denison, Konopczynski, Garcia, & Xu, 2006; Denison & Xu, 2010; Denison & Xu, 2014; Eckert, Call, & Rakoczy, 2017; Goldberg, 1966; Kushnir, Xu, & Wellman, 2010; Ma & Xu, 2011; Rakoczy et al., 2014; Téglás, Girotto, Gonzalez, & Bonatti, 2007; Téglás et al., 2011; Xu & Garcia, 2008; Yost, Siegel, & Andrews, 1962). For example, if a distribution has more red than white balls (e.g., 80 red and 20 white), they infer that a small sample taken from that distribution should also have more red than white balls. Although young children and non-human primates perform above chance on many probability problems, poor performance has been observed in some experiments, particularly in the 3- and 4-year-old age group (Girotto, Fontanari, Gonzalez, Vallortigara, & Blaye, 2016; Girotto & Gonzalez, 2008; Piaget & Inhelder, 1975). The current experiments explore whether some of this variability in performance is due to differences in problem difficulty by manipulating features of the problem that diagnose strategy use. We used a paradigm designed specifically for 3- and 4-year-olds to ensure that their abilities were not masked by difficulties with, or lack of engagement in, the task itself.

When adapting a task for a particular population, it is important to ensure that the paradigm is suitable to their abilities and still captures the essential aspects of the skill of interest. Issues regarding task-appropriateness have arisen throughout the course of research on children's probabilistic reasoning. Though Piaget's seminal work provides one of the first analyses of children's probabilistic reasoning abilities (Piaget & Inhelder, 1975), younger children's performance may have suffered due to the very high verbal demands of the task. Participants were asked which color item the experimenter was most likely to obtain on a random draw (Yost et al., 1962). Children's responses were then coded as correct based on their explicit reference of probabilistic concepts. From this work, it was concluded that children younger than 12 years of age struggled with probabilistic concepts. Conversely, presenting preschoolers with a choice paradigm suitable for infants and primates (e.g., Denison & Xu, 2014; Rakoczy et al., 2014) also appears to hinder their performance. When designing tasks for pre-verbal infants, experimenters use prompts that provide general encouragement (i.e., infants are told, "You can do it! Get the one you like!"). However, this prompt could make the task unclear to a preschooler with more advanced cognitive and linguistic abilities because these instructions are misleading. Children might recognize that when they choose something in a probabilistic context, they cannot guarantee that they will "get the one they want", they can only make a best guess. When this prompt was used with preschoolers, 3- and 4-year-olds' performance suffered (Girotto et al., 2016, Expt. 2). We used an age-appropriate method in the current experiments by asking children to provide a forced-choice response to a direct but simple probability question (see Procedure).

Moreover, there is considerable variability in the types of problems that have been presented to children in this age group. Falk, Yudilevich-Assouline, and Elstein (2012) outline this important point in their comprehensive assessment of school-aged children's probabilistic reasoning. Children were asked to choose between two small populations of items, each including a proportion of target and non-target items, to sample from in order to maximize their chances of obtaining a target item on a blind draw. The authors note that much previous research has overlooked the importance of manipulating numerical features of the presented problems when examining children's overall performance. Without manipulating these features across a variety of problems, it is difficult to know whether heuristic reasoning or true probabilistic inference led to correct responses in previous experiments. To combat this problem,

Falk et al. developed a battery of diagnostic problems that could not be solved using simple heuristic strategies (see Denison & Xu, 2014, for a similar approach with infants). For instance, in probability problems, children can use a heuristic in which they only compare the number of target items across populations, and thus ignore proportions. One can diagnose whether children are using this strategy by presenting them with problems that contain an equal number of target objects across two populations (e.g., 12 targets and 4 non-targets vs. 12 targets and 48 non-targets), and asking them to choose a population to draw from for the best chance of obtaining a target. This allows researchers to diagnose use of a strategy that solely focuses on choosing the population with more target objects, because children would be unable to solve such a problem if they tried to apply this strategy. One can also include problems in which there are more non-target objects in the more probable population to diagnose an avoidance strategy. Thus, children cannot succeed by simply choosing the population with more target items, or by choosing the population with fewer non-targets.

Notably, Falk et al. (2012) included problems in their experiment that assessed use of a good versus bad label shortcut. That is, instead of discerning the proportion of objects in each population, children could use a simpler shortcut that focuses on the majority type of objects in each population but does not require comparing proportions *across* populations. Many studies have presented children with a choice between, for example, a 75% target population and a 25% target population. A child could solve this problem by labelling the 75% population as "good", because the target objects are in the majority, and the 25% population as "bad", because the non-target objects are in the majority. This would lead them to approach the "good" population without carefully discerning and comparing the proportion of objects in each population. To assess use of this heuristic, Falk et al. included problems that were on the same side of ½. If a child who uses the good versus bad label shortcut was presented with two populations on the same side of ½, such as 75% and 95%, they would be unable to solve such a problem because both populations would receive the same label.

We attempted to tease apart preschoolers' use of true proportional reasoning from use of heuristics that approximate probabilistic inference. Because we were presenting these problems to children younger than those tested by Falk et al. (2012), we included problems that diagnosed use of simpler heuristics that may be used by preschoolers, as well as some of the more advanced ones described above. We included problems with more target objects in the less probable population and problems with an equal number of target objects in both populations. These features allowed us to examine if young children solely focus on target objects. Additionally, we included problems where the more probable population contained more non-target objects to examine if children attempted to avoid this option. We also included problems on the same side of ½ to gauge children's use of a shortcut that involves focusing on the majority of objects in individual populations.

Finally, closer, rather than more disparate, relative likelihoods (sometimes referred to as the "ratio of ratios") can make problems more difficult to evaluate. That is, when the likelihoods of each population are closer together, the problem can be more difficult to solve than when they are further apart because the populations themselves are more difficult to visually discriminate. For example, if Problem 1 contained a comparison between 80% and 75% targets, and Problem 2 contained a comparison between 90% and 60% targets, Problem 1 would be more difficult to solve because the relative likelihoods are closer and are more difficult to discriminate. Previous investigations of preschooler's probabilistic reasoning have not examined the impact of relative likelihood on their responses (but see Hoemann & Ross, 1971, for a similar manipulation using a spinner task), so we include a manipulation of this feature in the current experiments. Thus, for each problem type we included two versions, denoted 1 and 2, to mark, respectively, closer and further relative likelihoods.

Experiment 1

In Experiment 1, we presented 3- and 4-year-old children with a battery of probabilistic reasoning problems using a two-alternative forced choice procedure in a gumball machine paradigm. Children were tasked with selecting the population that was more likely to yield a blue object. We developed a set of problems to assess use of different strategies (see Figure 1). Problems A1 and A2 presented children with populations on the same side of ½. These problems also included more targets and more overall objects in the less probable population, so a child would not succeed on these problems if they were drawn to these features. Because this problem is challenging, the more probable population only contained target items, and thus the outcome was deterministic. Problems B1 and B2 were simple probabilistic comparisons that could be solved with multiple shortcuts. Although these simpler problems do not diagnose use of these shortcuts, we included them in our problem set to gauge the effectiveness of our paradigm with this age group, as 3- and 4-year-old children have solved these very simple problems in previous experiments. Problems C1 and C2 prevented children from selecting the population with more target objects, because the number of target objects was the same in both populations. Problem D presented children with two uniform populations in which one population only contained targets, and the other contained only non-targets to assess the effectiveness of the paradigm. This problem was always presented second to last, allowing us to gauge whether most children were following the task through such a large number of problems. Problem E was the inverse of Problem A1 and was included to diagnose whether children might use an avoidance strategy to solve problems (i.e., choosing a population that has fewer non-targets).

We included two versions of Problems A, B, and C to determine if probabilities that had higher relative likelihoods (i.e., problems that were further apart in probability, which were labeled with a 2), were easier for children to evaluate.

Methods

Participants Data from 50 3- and 4-year-olds were included in analyses (*mean age* = 4;2 [years;months]; *range* = 3;3 to 4;11). The sample size for the experiment was determined based on a power analysis for a larger study. An additional five children were tested and were excluded from analyses due to parental report of atypical development (n = 1), parental report of very low English exposure (i.e., hearing English less than 50% of the time; n = 2), and not finishing the task (n = 2). Participants were recruited from a database of families and received a small gift for their time.

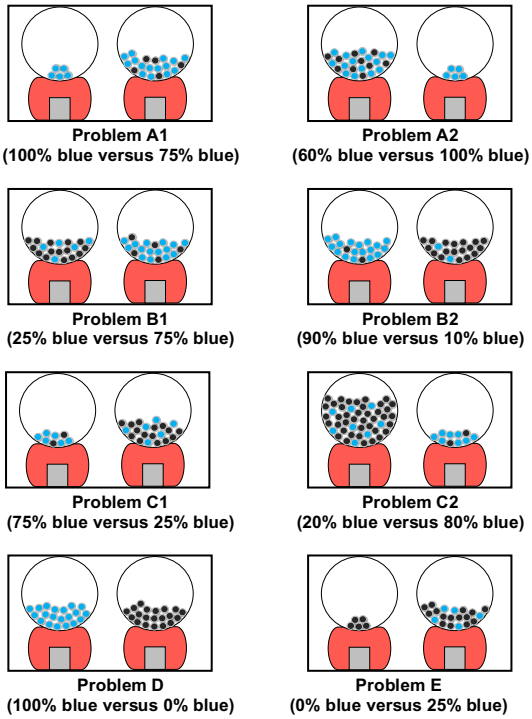

Figure 1: Probability problems presented in Experiment 1.

Materials and Procedure Participants were tested individually in a quiet room. The child and experimenter were seated together at a table. If the child's parent was present in the room, they were seated across the table from the child, unable to see the iPad's screen, and were asked to refrain from commenting or influencing their child's responses during the experiment. After the probability task, children completed measures assessing individual differences in their cognitive abilities (i.e., executive function and receptive vocabulary skills) as part of a larger study. Because the probability task was completed first, the additional measures had no impact on performance.

Probability problems were presented on an iPad, using a gumball machine paradigm. Prior to the test trials, the experimenter explained how gumball machines worked by showing participants two machines filled with a mixture of gumballs of various colors. To discourage children from focusing on the objects that were closer to the opening, the gumballs were then mixed and appeared in different positions in the machine, illustrating that any gumball in the population, regardless of its initial position, could be sampled. After mixing three times, each machine yielded one gumball. The experimenter told participants they had to choose between two machines and reiterated that each machine would only yield one gumball. Participants were told that they would receive a sticker if they chose a machine that yielded a blue gumball. Children then completed eight probability trials and were asked to choose the gumball machine that gave them the best chance of obtaining a blue gumball. On each trial, the machines always produced the more probable color gumball.

In populations that contained both colors, a blue and black gumball were positioned near the opening to ensure that children did not solely focus on the objects that were situated closer to the opening. The side of the correct gumball machine and the order each problem was presented were counterbalanced. Problems A, B, and C were counterbalanced in two blocks, with half of the participants completing version 1 in the first block. Problems D and E were always presented as problems 7 and 8, respectively. Problem D was presented second to last to so that we could assess whether children remained motivated throughout the task. Problem E was presented last; children were given a sticker for either choice, as black was the more likely outcome in both populations. Thus, it was presented last to ensure children did not expect to receive a sticker for a black gumball on subsequent problems.

Results and Discussion

Children received a score of 1 on each problem if they chose the machine that contained the higher proportion of blue (see Table 1 for means, standard deviations, and significance tests against chance for all problems).

Table 1: Children's performance in Experiment 1.

Problem	*M*	*SD*
A1	.82	.39
A2	.82	.39
B1	.90	.30
B2	.94	.23
C1	.82	.39
C2	.76	.43
D	.88	.32
E	.92	.27
Overall	.86	.17

Note: Individual problems were analyzed using binomial probabilities, overall score analyzed using single-sample t-test. All p values for the above analyses were ≤ .001.

We examined if children found some of the critical problem types (A through C) more difficult than others, and if they found problems with higher relative likelihoods easier to evaluate. To investigate this, we conducted a repeated-measures ANOVA with the critical problem types (A, B, C) and version (1, 2) as a within-subjects factor and child's age (younger half versus older half) as a between-subjects factor. There was a main effect of age, $F(1, 48) = 7.06$, $p = .01$, η^2_p

= .13, and problem type, $F(2, 96) = 4.15$, $p = .02$, $\eta^2_p = .08$, on children's scores. On average, the older children in the sample scored higher than the younger children (older children: $M = .91$, $SE = .04$; younger children: $M = .77$, $SE = .04$; $Mean_{Difference} = .14$, $p = .01$). Problem type B ($M = .92$, $SE = .02$) was significantly easier than problem type A ($M = .82$, $SE = .04$; $Mean_{Difference} = .10$, $p = .04$) and problem type C ($M = .79$, $SE = .05$; $Mean_{Difference} = .13$, $p = .01$). Problem version (i.e., relative likelihood) and all interactions were non-significant. Because problems D and E did not include these critical features and did not have a complement problem, we did not include them in these analyses. However, both problems were solved well-above chance (see Table 1). Performance on Problem D indicates that most children could still follow the task after they completed a number of more difficult problems. Moreover, the successful performance on Problem E suggests that children are not simply choosing the population with fewer non-targets.

To examine whether children's strong performance on probability problems was driven by learning over the course of the experiment (as the machines produced the more probable color on each problem type), we ran an additional repeated measures ANOVA with trial order (problem presented in place 1, 2, 3, 4, 5, 6) as a within-subjects factor. This analysis indicated that trial order did not significantly impact children's performance, $F(5, 240) = 1.22$, $p = .30$. Regardless of problem type, children performed well-above chance on trial 1 ($M = .90$, $SD = .3$, binomial, $p < .001$), which also suggests no effect of learning.

To summarize, Experiment 1 established that young children are able to solve probabilistic reasoning problems at rates well-above chance. Although the older children in our sample performed significantly better than the younger children, both age groups successfully solved the problems. Children performed significantly better on problem type B than types A and C, which is unsurprising due to the number of shortcuts they could have used to solve problems B1 and B2. Nevertheless, children still performed well on the more difficult problem types, suggesting that they do not solely rely on these heuristics.

Experiment 2

In Experiment 2, we presented a second group of children with more difficult problems to further test their use of various strategies. Because the children in Experiment 1 performed very well on our problems, we wanted to further explore their performance with a battery that contained some more challenging features (see Figure 2). Problems A1 and A2 presented children with two populations on the same side of ½, in which there were more targets and more overall objects in the less probable population. Problem types B and C presented children with two populations that had an equal number of target objects and more overall objects in the less probable population. Problems D1 and D2 contained more target objects and more overall objects in the less probable population. In this experiment, Problem E presented children with two uniform populations in which one population only contained blue gumballs, and the other contained only black (see Figure 1, Problem D). Because children in this experiment were presented with a more difficult set of problems, this problem was included again to gauge children's ability to follow the task. We included two versions of Problems A, B, C, and D to determine if probabilities that had higher relative likelihoods (i.e., problems that were further apart in probability, which were labeled with a 2), were easier for children to evaluate.

Methods

Participants Data from 74 3- and 4-year-olds were included in analyses (*mean age* = 4;2; *range* = 3;7 to 4;11). Again, the sample size was determined based on a power analysis for the larger study. An additional seven children were tested but were excluded from analyses due to parental report of atypical development ($n = 2$), parental report of very low English exposure (i.e., hearing English less than 50% of the time; $n = 2$), and not finishing the task ($n = 3$). Participants were recruited from a database of families and a daycare in the region. Children received a small gift for their time.

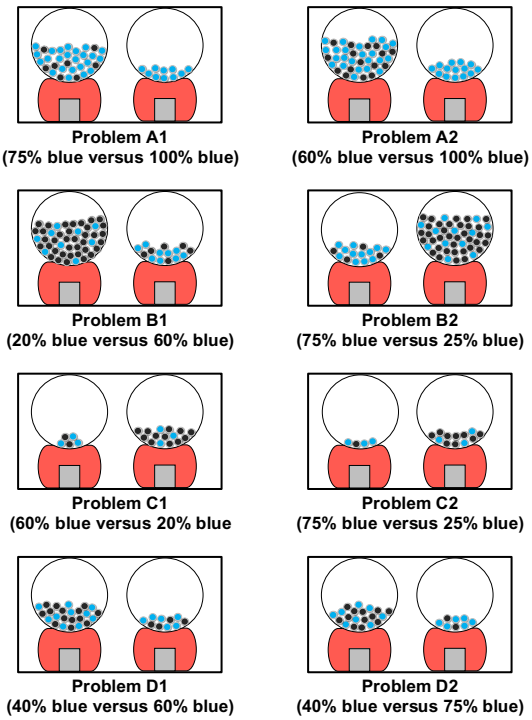

Figure 2: Probability problems presented in Experiment 2. *Note*: Problem E (not shown) was identical to Problem D in Experiment 1 *(see Figure 1)*

Materials and Procedure Participants were tested individually in a quiet room in the lab or at their daycare. The procedure was identical to Experiment 1, with the exception of the new battery of problems.

The probability problems were presented in the same manner as in Experiment 1. The side of the correct gumball machine and the order that each problem was presented were counterbalanced. Problems A, B, C, and D were

counterbalanced in two blocks, with half of the participants completing version 1 in the first block. To ensure that children remained motivated and followed the instructions throughout the task, Problem E was always presented last.

Results and Discussion

Children received a score of 1 on each problem if they chose the machine that contained the higher proportion of blue (see Table 2 for means, standard deviations, and significance tests against chance for all problems).

Table 2: Children's performance in Experiment 2.

Problem	*M*	*SD*
A1	.88	.33
A2	.88	.33
B1	.64	.49
B2	.76	.43
C1	.74	.44
C2	.84	.37
D1	.73	.45
D2	.78	.41
E	.93	.25
Overall	.80	.17

Note: Individual problems were analyzed using binomial probabilities, overall score analyzed using single-sample t-test. All p values for the above analyses were ≤ .001, with the exception of B1 (p = .03).

Similar to Experiment 1, we explored if children found some of the critical problem types (A through D) more difficult than others, and if they found problems with higher relative likelihoods easier to evaluate. To examine this, we conducted a repeated-measures ANOVA with the critical problem types (A, B, C, D) and version (1, 2) as a within-subjects factor and child's age (younger half versus older half) as a between-subjects factor. There was a main effect of problem type, $F(3, 216) = 5.36$, $p = .001$, $\eta^2_p = .07$, and version, $F(1, 72) = 4.65$, $p = .03$, $\eta^2_p = .06$, on children's scores. Problem type A ($M = .88$, $SE = .03$) was significantly easier than problem type B ($M = .70$, $SE = .04$; $Mean_{Difference} = .18$, $p < .001$) and problem type D ($M = .76$, $SE = .04$; $Mean_{Difference} = .12$, $p = .007$). Problem type C ($M = .79$, $SE = .03$) was marginally more difficult than problem type A ($Mean_{Difference} = -.09$, $p = .07$) and marginally easier than problem type B ($Mean_{Difference} = .10$, $p = .06$). Problems labeled with 2 ($M = .81$, $SE = .03$), comparisons that were further apart in relative likelihood, were significantly easier than problems labeled with 1 ($M = .75$, $SE = .03$; $Mean_{Difference} = -.07$, $p = .03$). Children's age and all interactions were non-significant. Because Problem E did not include these critical features and did not have a complement problem, it was not included in this analysis. However, as seen in Table 2, this problem was again solved well-above chance, indicating that most children still followed the task after they completed a number of more difficult problems.

To examine learning over the course of the experiment, we ran an additional repeated measures ANOVA that included counterbalanced trial order (problem presented in place 1, 2, 3, 4, 5, 6, 7, 8) as a within-subjects factor. There was an effect of trial order, $F(7, 511) = 2.31$, $p = .03$, $\eta^2_p = .03$, on children's scores, with scores improving over the session. Though this effect of order suggests that some learning may have occurred throughout the experiment, children performed well above chance on trial 1 ($M = .77$, $SD = .42$, binomial $p < .001$), indicating that learning did not entirely account for the strong performance.

In Experiment 2, we presented children with more challenging probabilistic reasoning problems. Although they were presented with this more difficult battery, children still performed at rates well-above chance across the age group. Problem type A was relatively easy for participants to solve, possibly because the correct option only contained target gumballs. Compared to the other problems, children found problem types B and D more difficult. On those problems, children were unable to rely on a number of cues, including the number of targets and the number of overall objects. We also found that children performed better on problems labelled with 2, which had relative likelihoods that were further apart and thus were easier to visually discriminate. Finally, although children performed above chance on the first trial, we observed an effect of trial order on performance, suggesting that learning may have contributed to performance.

General Discussion

In two experiments, we established that 3- and 4-year-old children are able to reason about probabilities at rates well-above chance. Though the older children in our sample performed significantly better than the younger children in Experiment 1, we did not find any age differences in performance in Experiment 2. Problems that contained multiple shortcuts or a deterministic outcome were easier for children to solve, and relative likelihoods impacted performance in Experiment 2 with our more difficult set of problems. Though children in both experiments performed above chance on the first trial, feedback may have affected children's scores over the course of Experiment 2.

Differences between our design and those of previous experiments may have facilitated performance in our paradigm. Children in our experiments were asked an age-appropriate question about probability. The verbal demands of the task affected preschoolers' performance in the past, as their performance suffered in paradigms with very high and very low verbal demands. In contrast to using verbal explanations as a dependent measure (as in Piaget & Inhelder, 1975), or using verbal cues that might have been too general (as in Girotto et al., 2016), children provided a forced-choice response to a simple, explicit question about probability. This method appears to have suited their abilities.

Children may have also found our gumball machine paradigm, which was presented on an iPad, engaging, and this may have helped maintain their interest over a number of trials. This design allowed us to display the contents of the gumball machine clearly, and the objects remained in view while the child made their choice. In some previous experiments, the experimenter sampled a hidden object from

each population and would ask the child to choose between the two hidden samples. Displaying the populations during the child's choice may have eliminated a working memory demand, because children did not have to maintain a representation of the populations during the sampling process. To disentangle the influence of these features, future work could again present children with two gumball machines on an iPad, though the populations would be covered while a hidden object is drawn from each machine. This would help us determine if clearly displaying the objects aids performance, and if hiding the objects during the sampling procedure creates a working memory demand.

We also provided children with feedback for their performance on each trial, and they were shown the most probable outcome from both populations after they made their choice. We used feedback to help sustain motivation over the course of the experiment because we were presenting very young children with multiple trials. Although we found no evidence of learning in Experiment 1, we found an effect of trial order on performance in Experiment 2. Nevertheless, children in both experiments performed above chance on the first trial prior to receiving any feedback. To further investigate learning in this context, future work could test the effectiveness of feedback at combating the use of overlearned strategies that approximate probabilistic inference. In turn, this work would shed light on how more sophisticated probabilistic reasoning strategies are acquired and fine-tuned with practice.

Moreover, the current experiments explored various strategies that children could use to approximate probabilistic inference. Though older children are drawn to populations with more target objects (i.e., denominator neglect; Falk et al., 2012), preschoolers in our experiments performed well on problems in which the less probable population contained more target objects, and when the number of target objects were equated. One notable difference between our problems and those that older children struggled with is that older children are typically presented with more difficult problems, in which the relative likelihoods are more difficult to discriminate. In our problems, the relative likelihoods were more distinct, making the problems easier overall. Surprisingly, preschoolers were drawn to populations with more *overall* objects (that is, target plus non-target). In both experiments, children's performance was slightly worse on problems where the less probable population noticeably contained more objects. Though older children are not drawn to populations with more objects (Falk et al., 2012), the current findings suggest that the overall number of objects is a salient feature for preschoolers. Because of this somewhat surprising finding, we are currently developing a battery of problems to further clarify how features of the problem, such as overall objects and number of targets, affect young children's probabilistic reasoning performance. Future work with a larger age range could also investigate how use of different strategies varies over the course of development.

We presented preschoolers with problems that were on the same side of ½ to explore nuances in their ability to compare proportions. Children in both experiments were able to make these comparisons and considered the proportion of objects in each population, even though the less probable population contained more target objects. Because we were unsure if preschoolers could solve these more difficult problems, the more probable population was uniform and only contained target objects. Inclusion of the uniform population allowed for a straightforward assessment of children's reasoning abilities, serving as a first step in pitting true proportional reasoning against a heuristic that focuses on the absolute number of target objects. Although this first step established that they are able to make these comparisons, future work should present preschoolers with two *probabilistic* populations (i.e., both contain target and non-target objects) on the same side of ½. This comparison is more difficult, because children are comparing two probabilistic populations and, by the nature of this design, the relative likelihoods are closer together. Though relative likelihood did not influence performance on Experiment 1, it impacted preschoolers' responses on the more difficult battery in Experiment 2. Thus, future work should continue to test the impact of relative likelihood on preschooler's performance, notably in cases where both populations are on the same side of ½.

Finally, we used two sets of problems to assess preschoolers' probabilistic reasoning in the current experiments. Though both batteries indicated that children could successfully reason about probability, one may wonder which battery would best provide an overall assessment of a child's abilities. For space and intended focus of the current paper, we did not report the results of a large set of individual difference measures that were collected with the children that assessed their executive function and receptive vocabulary abilities. These measures tend to correlate well with children's quantitative and general reasoning abilities during early childhood. However, the battery used in Experiment 1 correlated well with these measures, while the battery in Experiment 2 did not show as strong of a relationship. Therefore, at the present time, the problems in Experiment 1 might be the best set to use when gauging children's abilities in future work. We are currently working on another battery of problems, which include some problems from Experiments 1 and 2, and some additional problems of even greater difficulty to continue refining the set.

In sum, preschool children in both experiments solved probabilistic reasoning problems at rates above chance. The current findings illustrate the importance of using an age-appropriate paradigm when establishing the abilities of a particular population. Though children did not rely solely on erroneous strategies, future work is needed to explore how features of probabilistic problems impact performance.

Acknowledgments

These experiments were conceived of in collaboration with Tara McAuley and Bethany Nightingale as part of a larger project and we thank them for their insights. We thank members of the Developmental Learning Lab for help with data collection. Special thanks to parents and children for

participating. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to S. D.

References

Denison, S., Konopczynski, K., Garcia, V., & Xu, F. (2006). Probabilistic reasoning in preschoolers: Random sampling and base rate. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1216–1221).

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.

Eckert, J., Rakoczy, H., & Call, J. (2017). Are great apes able to reason from multi-item samples to populations of food items? *American Journal of Primatology, 79*(10).

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*(2), 207-233.

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,* 325 – 344.

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

Hoemann, H. W., & Ross, B. M. (1971). Children's understanding of probability concepts. *Child Development*, 221-236.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of others. *Psychological Science, 21*(8), 1134–1140.

Ma, L., & Xu, F. (2011). Young children's use of statistical sampling evidence to infer the subjectivity of preferences. *Cognition, 120*(3), 403-411.

Piaget, J., & Inhelder, B. (1975). *The origins of the idea of chance in children.* New York, NY: Norton & Company.

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

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

Yost, P. A., Siegel, A. E., & Andrews, J. M. (1962). Nonverbal probability judgments by young children. *Child Development, 33*(4), 769-780.