Charting children's fruit categories with Markov-Chain Monte Carlo with People

Pablo León-Villagrá[1] **Isaac Ehrlich**[2] **Christopher G. Lucas**[3] **Daphna Buchsbaum**[1]
[1]Brown University, USA [2]University of Toronto, Canada [3]University of Edinburgh, UK

Abstract

Uncovering how categories develop through childhood is crucial for cognitive science. However, even for simple domains, categories can be complex, making it challenging to access them experimentally, especially in developmental studies. Markov-Chain Monte Carlo with People ($MCMC_p$) is a statistically-based procedure that allows us to elicit category members from participants' implicit categories. However, due to the complexity of the paradigm, $MCMC_p$ has been limited to experiments with adult populations. Here, we develop and validate a child-friendly method for applying $MCMC_p$, producing the first $MCMC_p$ experiment to elicit category examples from children. Comparing fruit category members for five-year-olds and seven-year-olds, we find generally consistent representative fruits and a developmental progression of initially broad and overlapping fruit categories to more differentiated distributions.

Keywords: Categorization; cognitive development; experimental methods

Introduction

Uncovering how we represent categories and how these representations change throughout development and learning is crucial for cognitive psychology. Therefore, sophisticated experimental methods have been developed to determine properties of these mental representations, such as using similarity judgments, spatial sorting, or generalization gradients to determine psychological spaces (Shepard, 1980, 1987; Goldstone, 1994; Hout, Goldinger, & Ferguson, 2013).

However, previous research on conceptual development has typically used small sets of hand-picked, simplified items (e.g., stylized images of animals) to keep the experimental duration manageable and to ensure that children recognize the presented items. These materials restrict the insights studies can offer about the structure of children's categories. This is because the items an adult selects as salient or representative category members might not correspond to what a young child considers representative, introducing a potential bias into the experiments. In addition, many experiments assess children's categories by presenting the child with discrete choices between contrasting materials. For example, in feature generalization tasks, it is common to ask the child to generalize the property of an example category member to one of two options, where each option usually matches the example category in some way (e.g., a visually similar member of a different category versus a visually distinct member of the same category, where adult experimenters determine similarity and category membership). These discrete choices can reveal if an example is perceived to be a category member or inferred to have a particular property. However, category membership is graded, with some instances being judged better category members than others (Rosch & Mervis, 1975), and atypical instances can lie at the boundary of competing categories, for example, when deciding what kind of animal a platypus is, or what color aquamarine is. Reducing the number and variety of materials used in experimental studies and testing discrete, either-or category membership thus reduces our ability to answer more fine-grained questions, such as how representative category members develop and how category boundaries are formed.

Here we adopt an experimental paradigm, Markov-Chain Monte Carlo with people ($MCMC_p$; Sanborn, Griffiths, & Shiffrin, 2010), that is not limited to testing a pre-specified set of materials. Instead, the method adaptively learns to present representative stimuli to participants and then explores participants' categories, presenting stimuli proportionally to the stimulus' degree of category membership.

For example, when eliciting repeated examples of what an apple is, there might be many Red Delicious, some but fewer Granny Smiths, more red and green apples than yellows, and few, if any, Black Oxfords. $MCMC_p$ asks participants to repeatedly choose between a previously selected category member, in our example, an apple, and an adaptively generated example of the category (the proposed update). Participants are asked to choose the more likely category member of the two, e.g., "*Which is the most apple-like?*". If this new proposal is selected, it becomes the next example, or else the previous state is carried forward.

$MCMC_p$ is motivated by Markov-Chain Monte Carlo (MCMC), a widely used statistical method. In many inference problems, it is impossible to directly calculate measures of a variable of interest, such as the mean or variance. MCMC allows practitioners to approximate these measures by producing many samples that approximate the distribution of interest.

MCMC produces these samples by iteratively extending a chain of states, each state corresponding to one sample from the target distribution. Many MCMC algorithms exist; here, we will focus on one of the most common variants, Metropolis–Hastings MCMC (MH-MCMC). In each iteration, MH-MCMC generates a proposed update by perturbing the previous state. Then, the likelihood of the proposed update and the previous state are compared. If the update is accepted, the proposal becomes the new state in the MH-MCMC chain,

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

otherwise, the previous state is maintained. It can be shown that for appropriate proposal and acceptance procedures, MH-MCMC produces a chain of samples approximating the desired probability distribution; for an accessible introduction to MCMC, see MacKay (2003).

In MCMC_p, the desired probability distribution is the participants' distribution over category members. As in MCMC, we cannot access this distribution directly, but we can present participants with a series of forced choices between two category instances, where one option is the previous choice and the alternative is a proposed update. Given that these forced choices correspond to a valid acceptance procedure for MCMC (Sanborn et al., 2010), the sequence of choices will approximate the participant's category member distribution.

MCMC_p offers several advantages over alternative experimental methods. Most importantly, in MCMC_p, the experimenter does not need to specify all stimuli before running the experiment. Instead, stimuli are generated adaptively, allowing experiments to test complex stimuli effectively and reducing bias introduced by pre-selecting materials. MCMC_p offers unique potential, especially in studies where the experimenter's intuitions about which stimuli to test might be misaligned with the participant population, such as in developmental studies.

MCMC_p was first introduced by Sanborn et al. (2010) and has since been used to produce category representations for complex and diverse categories such as stylized fruits and animals (Sanborn et al., 2010), face perception (McDuff, 2010; Martin, Griffiths, & Sanborn, 2012), or continuous-valued relationships (León-Villagrá, Klar, Sanborn, & Lucas, 2019). However, MCMC_p typically requires hundreds or thousands of samples to capture the structure of a category (Sanborn et al., 2010; McDuff, 2010), resulting in repetitive and long experimental sessions. Therefore, this approach has been limited to adult populations.

Here we test, for the first time, if MCMC_p can be used in developmental experiments. To make experimental sessions more manageable for young children, we use short MCMC_p chains, using the final state of a previous participant as their starting point (Martin et al., 2012; Ramlee, Sanborn, & Tang, 2017; León-Villagrá, Otsubo, Lucas, & Buchsbaum, 2020). We have previously shown that these linked MCMC_p sessions can produce population-level category members that are qualitatively similar to traditional MCMC_p designs (León-Villagrá et al., 2020). Furthermore, we introduce a child-friendly cover story to make the task more engaging for children and motivate them to select representative category members. We show that the resulting experimental setup allows us to efficiently produce examples of children's categories: MCMC_p quickly learns to produce representative category members and then explores the extent of the categories. Uncovering children's categories efficiently offers the prospect of answering fundamental questions in cognitive development and cognitive science, such as if children's conceptual spaces undergo sudden restructuring or are, instead, gradually refined.

Experiment

We assess children's category representations for three fruit categories, apples, oranges, and grapes. We chose this domain since previous adult MCMC_p experiments have produced convincing category representations (Sanborn et al., 2010; León-Villagrá et al., 2020), and previous experiments showed that children were comfortable treating the stylized experimental materials as fruit-like (León-Villagrá, Ehrlich, Lucas, & Buchsbaum, 2022). In pilot experiments, we observed that children would quickly realize that they could select fruits freely without "erroneous" choices affecting their performance. As a result, several children in our pilots enjoyed selecting implausible fruits, such as blue oranges. To motivate children to select representative fruits, we thus introduced a cover story in which children were instructed that their MCMC_p choices served to teach a robot to draw fruits. Introducing this cover story discouraged purposefully selecting nonsensical fruit examples and was an important modification to make MCMC_p practicable as a developmental paradigm.

Participants

Following our pre-registered participant collection criteria,[1] a total of 99 children participated in the task, split into two age groups, 5-year-olds ($N = 47$) and 7-year-olds ($N = 52$).[2] Children were recruited from the lab's parent databases, the online volunteer database children helping science, and Facebook parent groups. Note that our collection criteria did not specify a fixed number of participants. Instead, we aimed to select a minimum number of participants required for MCMC_p to showcase its practical use as a developmental paradigm. Therefore, we pre-registered the number of trials required to achieve reliable MCMC_p results based on statistical criteria (see Results) and collected data until these criteria were achieved. Most children spent less than 10 minutes on the MCMC_p task ($M = 8.2$ minutes, $SD = 2.2$), and the total experiment (including briefing, familiarization, follow-ups, and debriefing) took about 25 minutes.

Additionally, 49 5-year-olds and 21 7-year-olds participated in the task but were excluded according to our pre-registered exclusion criteria. Most exclusions were due to inattention ($n_5 = 21$, $n_7 = 7$), random clicking ($n_5 = 21$, $n_7 = 4$) or technical issues ($n_5 = 4$, $n_7 = 7$).

[1] `https://osf.io/vnkaf`

[2] We originally planned to compare 4-5 versus 6-7-year-olds. However, early in data collection and prior to any analysis, we noticed that our pre-registered exclusion criteria resulted in high exclusion rates for 4-year-olds. After testing 41 children, 9 of 12 4-year-olds were excluded due to inattention or random clicking. In contrast, only 1 of 8 5-year-olds, 1 of 9 6-year-olds, and 2 of 12 7-year-olds were excluded. To allow for balanced age groups while keeping the participant numbers manageable, we dropped 4- and 6-year-olds from the design and continued with only 5- and 7-year-olds. Therefore, a small number of 4 and 6-year-olds are included, but exclusively in the early stages of burn-in, and thus do not affect the interpretation of the resulting fruit category member distributions.

Materials

The experiment was conducted in a web application on the participants' computers or tablets. The stimuli were stylized fruits, as introduced originally in Sanborn et al. (2010). These fruits were programmatically created by varying the radii of three circles (r), as well as the horizontal (h) and vertical (v) distance between the circles. To create a shape, we calculated the convex hull over the three circles and programmatically colored the shape (variables *hue*, *saturation*, *luminosity*). Each fruit was topped with a brown stem to indicate the fruit's orientation to the participants. By varying these six parameters, complex colored shapes could be created; see Figure 1.

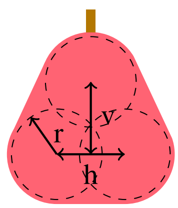

Figure 1: Stimuli were programmatically generated by specifying radii (r), vertical (v), and horizontal (h) length, and three color parameters.

Procedure

The experiment consisted of four blocks and was conducted over Zoom video conferencing. Participants were linked to previous (non-excluded) participants within their age group, resulting in an overall linked sequence of category samples representative of all children in the age group.

Setup and Familiarization: The first block of the experiment consisted of an audio and video setup and briefing (with the child's guardian present), familiarizing the child with the task, and testing their comprehension. To motivate children to select representative fruit examples, we presented a short video that provided the rationale for the MCMC$_p$ task. In the video, children were introduced to Robbie, a smart robot that does not know how to draw fruits. After the video was presented, the children were asked if they could help Robbie learn how to draw fruits. Once they agreed, they performed one test trial, in which they had to select a banana from two options ("*Robbie drew these. Can you pick the banana?*"). In the test trial, one stylized banana and a round blue shape were presented in the same fashion as the later MCMC$_p$ trials. For an illustration of the familiarization, see Figure 2.

MCMC$_p$ Blocks: The main experiment used a custom web application. During the child's interaction, the experimenter recorded the child's engagement with the app and provided feedback and encouragement when required. The main experiment consisted of two blocks of MCMC$_p$, each with 54 trials. In each trial, an animation of Robbie drawing two fruits on two stylized white plates equidistant from the center of a

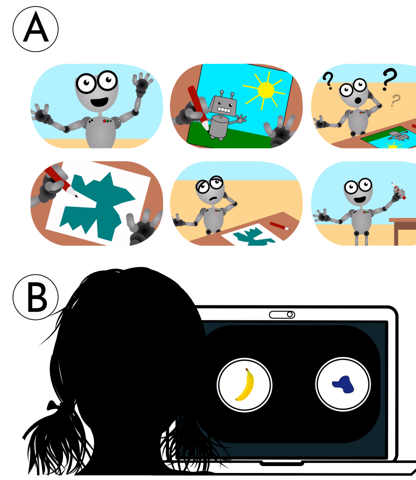

Figure 2: In the familiarization block, children were first shown a short video about the task context (A). In the video, Robbie the robot is introduced as a great artist who does not know how to draw fruits. The child is tasked to teach Robbie how to draw fruits. The second part of the familiarization consisted of one test trial, in which children were asked to select the banana from two options (B).

black screen was presented. This animation was followed by pre-recorded audio instructions, asking the child to "*Pick the [fruit]*" for one of the three fruit categories: apple, orange, or grape; for an illustration of the main experiment, see Figure 3.

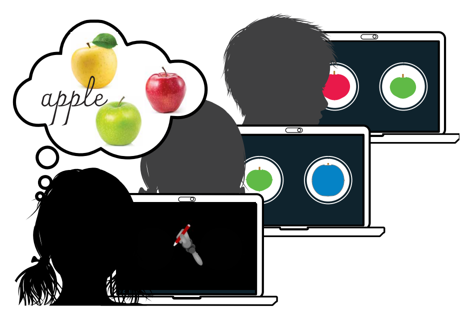

Figure 3: Illustration of the MCMC$_p$ procedure. In each trial, an animation showed Robbie "drawing" two fruits. Children were then asked to "*Pick the [fruit]*", for apples, oranges, and grapes. After maximally two blocks of 54 trials, the current selections for all three fruit categories were linked to the subsequent participant, thereby linking individual MCMC$_p$ chains.

In each trial, one fruit was the current state of the MCMC chain, and the other was the proposed update, with sides presented in random order. Children could select one of the two options using their mouse or touch. Upon selection, a sound was played, and an animation removed the two fruits and plates before the procedure was repeated with the next iteration of choices. To provide feedback on the child's progress, after each trial, a star appeared at the top of the screen. After 54 trials, a new screen appeared with Robbie presenting "*All the fruits he had drawn in his notebook*". Children could take a short break before continuing to the second block of 54 trials. If the child decided to end the task, or once both blocks of 54 trials were completed, the child proceeded to the final block.

To make the iterative procedure less obvious to participants, fruits were presented in an interleaved fashion, starting with apples. To assess whether $\mathrm{MCMC_p}$ produced representative fruit category members using established statistical measures from the MCMC literature (see Results section), we also interleaved three $\mathrm{MCMC_p}$ chains per fruit. Therefore the sequence of presentations was: $apple_1$, $orange_1$, $grape_1$, $apple_2$, etc.

To improve the speed at which $\mathrm{MCMC_p}$ moved toward representative category members, the first four $\mathrm{MCMC_p}$ blocks frequently generated proposals using uniform distributions over the shape or color parameters (three equally probable proposal schemes: uniform proposals in color, uniform proposals in shape, and Gaussian proposals centered on the previous fruit parameters). In contrast, later blocks favored Gaussian proposals (8/10 Gaussian proposals, 1/10 uniform in color, 1/10 uniform in shape). The standard deviation of the Gaussian proposals was set to cover 7% of each parameter range, as in Sanborn et al. (2010). As in Sanborn et al. (2010), proposals outside the valid parameter ranges were automatically rejected without being shown to participants; the current state was recorded, and the proposal counted as a rejection. A new proposal was generated until the parameters were within the admissible range.

Follow-up Questions: In the final block, children were asked follow-up questions about how they selected the three fruits ("*How did you pick the [fruit]?*", for apples, oranges, and grapes). Then they were asked, "*Did Robbie get better at drawing fruits?*", "*Did you enjoy this game?*", and if they could hear and see the task and experimenter clearly. Finally, the children were thanked and presented with their participation certificates.

Results

Determining the effectiveness of $\mathrm{MCMC_p}$

To determine if $\mathrm{MCMC_p}$ is a valid developmental paradigm, it is important to assess if $\mathrm{MCMC_p}$ produced consistent estimates of participants' categories. This analysis follows the common practice in MCMC of determining how consistently the multiple MCMC chains (in our case, three for each fruit) move across the parameter space. Many diagnostics have been proposed in the statistical literature; here, we focus on the rate of convergence, the effective sample size (the number of uncorrelated $\mathrm{MCMC_p}$ samples), the number of proposals that participants accepted, and the degree to which $\mathrm{MCMC_p}$ explored the parameter space.

Burn-in: It is common practice to start parallel MCMC chains at random locations in the category space to assess at which point these initially separate chains converge to an area of the category space where the distribution is concentrated. Before that point, MCMC samples will be biased by the random starting positions and will not accurately reflect the target distribution. Thus, it is standard practice to remove iterations before this point, the so-called burn-in period.

In $\mathrm{MCMC_p}$, determining burn-in is an important measure of the validity and efficacy of the method. $\mathrm{MCMC_p}$ would not produce interpretable category member distributions if the randomly initialized chains do not converge to the participant's category distribution, or if convergence requires impracticable numbers of iterations.

Thus, we determined the length of burn-in following preregistered criteria, by incrementally calculating the multivariate scale reduction ($\hat{R}$; Brooks & Gelman, 1998) of each fruit. After each participant, we calculated $\hat{R}$ values for all three fruits from the beginning of the chain (the first iteration in the sequence of linked $\mathrm{MCMC_p}$ trials) to iterations up to the current participants' last trial. Once a sequence was found that suggested that the chain had converged ($\hat{R}$ values ≤ 1.5, as specified in our pre-registration) we labeled this point as the end of the burn-in period. Both 7-year-olds ($M = 170$, $SD = 120.50$) and 5-year-olds ($M = 75.33$, $SD = 4.04$) converged within the expected number of trials, suggesting that $\mathrm{MCMC_p}$ successfully converges to children's category member distributions.

Since we obtained three different $\hat{R}$ values, one per fruit category, we conservatively used the maximum iteration across all three fruits; for the burn-in points per fruit and age group, see Table 1. We analyze category distributions after burn-in up to 1500 iterations.

Table 1: The iteration index i at which burn-in was achieved after incrementally increasing the subsequence i to j that resulted in $\hat{R} \leq 1.5$. Since most children performed two blocks of $\mathrm{MCMC_p}$, burn-in was achieved after 8-18 participants.

		i	j	Block	$\hat{R}$
5-years	Apple	80	288	16	1.46
	Orange	73	360	20	1.12
	Grape	73	324	18	1.49
7-years	Apple	309	648	36	1.21
	Orange	106	594	33	1.18
	Grape	95	594	33	1.13

Effective Sample Size: In addition to converging to the target fruit distribution, we also require sufficiently many samples to obtain reliable estimates of children's category member distributions. Since each state in a chain of MCMC or $\mathrm{MCMC_p}$ states depends on the previous, samples are correlated. Thus, the total iteration length (1500) tends to significantly overestimate the effective number of independent (uncorrelated) samples. A common way of estimating these independent samples is to estimate the effective sample size (ESS Gelman et al., 2013). Both 5-year-olds ($M = 43.3$, $SD = 21.56$) and

7-year-olds ($M = 29.84$, $SD = 17.18$) produced satisfactory per-parameter ESS values (for each fruit) and overall comparable ESS to previous adult experiments (Sanborn et al., 2010; León-Villagrá et al., 2020).

Acceptances: Another important measure to assess the effectiveness of $MCMC_p$ is considering the quality of proposals generated. Ideally, proposed states in MCMC strike a balance between presenting highly representative samples and exploring less representative areas of the category space (to ensure not missing potentially more representative areas).

Since out-of-range parameters were automatically rejected, the number of total samples was often higher than the number of trials seen by participants. Therefore, we evaluated acceptances, including automatic rejections, as they are diagnostic for the sampling process (total acceptance rates) and excluding automatic rejections (human acceptance rates), as this is diagnostic for the psychological validity of the method, see Table 2.

Table 2: Acceptance rates including automatic rejections (*acc*), and excluding automatic rejections (*hac*).

		M_{aac}	SD_{aac}	M_{hac}	SD_{hac}
5-years	Apple	0.33	0.47	0.38	0.49
	Orange	0.42	0.49	0.43	0.50
	Grape	0.42	0.49	0.41	0.49
7-years	Apple	0.23	0.42	0.24	0.43
	Orange	0.25	0.43	0.25	0.43
	Grape	0.27	0.44	0.27	0.44

Including automatic rejections, 5-year-olds ($M = .39$, $SD = .49$) and 7-year-olds ($M = .25$, $SD = .43$) produced satisfactory acceptance rates. Excluding automatic rejections, 5-year-olds ($M = .41$, $SD = .49$) and 7-year-olds ($M = .25$, $SD = .43$) produced very similar acceptance rates, suggesting that participants only rarely moved towards the fringes of the parameter ranges. Across age groups, these acceptance rates were higher than previously reported results in Sanborn et al. (2010) and León-Villagrá et al. (2020) and comparable to the recommended 20 % to 40 %(Roberts, Gelman, & Gilks, 1997).

Overall, these acceptance rates suggest that the proposal schemes were highly effective in producing representative fruits for the three fruit categories.

Multivariate Scale Reduction Finally, since participants could start diverging, even after burn-in was achieved[3] we determine $\hat{R}$ for the 1500 samples after burn-in. Overall $\hat{R}$ was close to the recommended convergence measures in the statistical literature ($\hat{R} \approx 1.1$) for all fruits, suggesting that $MCMC_p$ produced samples from the participants' category representations. Similar to the $\hat{R}$ values that determined burn-in, younger children had smaller $\hat{R}$ values ($M = 1.12$, $SD = 0.02$) than 7-year-olds ($M = 1.26$, $SD = 0.15$), and both age groups produced highly satisfactory convergence measures. These results suggest that $MCMC_p$ effectively converged to the category member distributions, and, within age groups, participants agreed on which fruit category members were representative.

[3]For example, if subpopulations of participants have significantly different category member distributions and these subpopulations cluster within the linked $MCMC_p$ chain.

Fruit Category Distributions

Since we have established that the $MCMC_p$ samples of both age groups converged to the fruit categories and that both groups generated sufficient samples, we next discuss these distributions further.

Visually, we found that 5-year-olds produced broader category distributions across parameters, and variances were more than twice as large as for 7-year-olds. Disregarding the broader distributions, 5-year-olds produced fairly consistent fruit distributions to 7-year-olds — both age groups agreed on which fruits had multiple characteristic colors (grapes) and that oranges only had one characteristic color, as well as which features were characteristic of fruits. For example, both age groups selected blueish and greenish grapes, orange oranges, and reddish and greenish apples; see Figure 4 for fruit means, and Figure 5 for histograms.

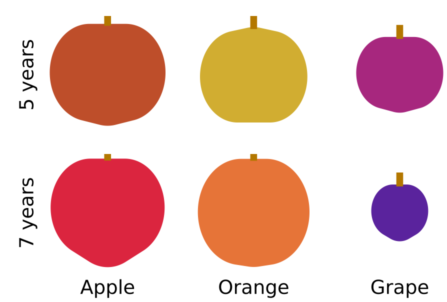

Figure 4: The mean fruit category members for 5- and 7-year-olds.

However, this visual comparison also suggests differences. Interestingly, for grapes, 5-year-olds produced considerably larger grapes (see the distribution of h, v, r parameters) than 7-year-olds. Finally, for 5-year-olds, the fruit parameters of the three fruits overlapped considerably, while 7-year-olds produced more differentiated categories.

Qualitative Features & Follow-up Questions

To assess how children chose the fruits, we additionally analyzed the reasons they provided for their choices after the $MCMC_p$ blocks. Responses were broadly consistent with the posterior category member distributions. Both 5-year-olds and 7-year-olds frequently named color and shape as a reason for their choices. Consistent with the posterior distributions, 7-year-olds mentioned *size* at higher rates than 5-year-olds when picking grapes, and both age groups mentioned *size* less for apples and oranges. Finally, older children mentioned *because*

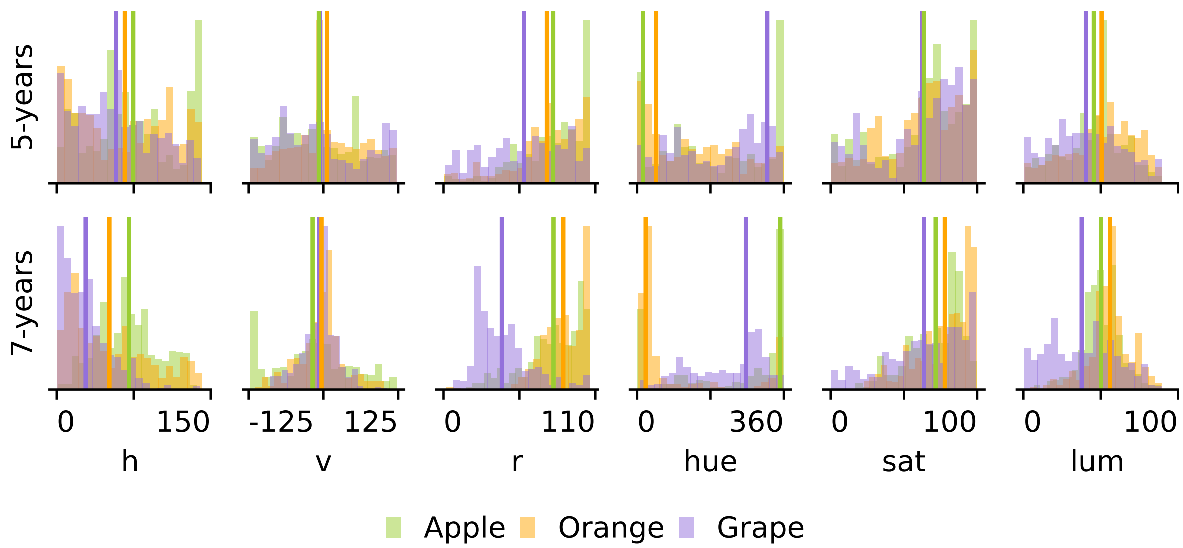

Figure 5: The category member distributions for 5-year-olds (top row) and 7-year-olds (bottom row). Each column displays the posterior histograms for one of the six parameters that determined the shape and color of the fruits for the three fruits. Note that hue is structured in circular space, with values wrapping around at 360°. Vertical lines show (circular) means.

it has a stem more often than younger children, suggesting that some children judged the materials as actual fruits, see Table 3.

Table 3: Proportion of children naming each feature when prompted "*How did you pick the [fruit]*". *Is* responses were statements such as "*Because it is a [fruit].*".

		Color	Shape	Size	Stem	Is
5-years	Apples	0.60	0.40	0.17	0.00	0.03
	Oranges	0.63	0.43	0.14	0.00	0.06
	Grapes	0.57	0.31	0.31	0.03	0.06
7-years	Apples	0.83	0.67	0.10	0.10	0.02
	Oranges	0.67	0.71	0.10	0.07	0.05
	Grapes	0.60	0.45	0.50	0.00	0.10

Most children reported that they enjoyed the game (93% of 5-year-olds, 94% of 7-year-olds) and saw and heard the task and experimenter clearly (100% of 5-year-olds, 96% of 7-year-olds). Overall, children perceived the drawings to improve (89% of 5-year-olds, 94% of 7-year-olds). Together with our convergence results, these findings suggest that we succeeded in adapting $MCMC_p$ as a developmental paradigm.

Discussion

We developed and validated a child-friendly method for producing category member distributions from young children without pre-specifying experimental materials. Using rigorous measures from the statistical MCMC literature, we found that $MCMC_p$ converged to children's categories and effectively produced samples from the three fruit distributions. Across measures, and for both age groups, we obtained similar, or better, statistical measures than previous $MCMC_p$ experiments with adults, suggesting that $MCMC_p$ can be used in developmental paradigms. By linking individual $MCMC_p$ runs together and introducing a child-friendly cover story, we showed that $MCMC_p$ can be used as a developmental paradigm, even in remotely deployed experiments. Analyzing the distributions over fruit category members, we found broadly consistent representative fruits and several categories characterized by multiple modes. However, we also found characteristic differences – younger children's grape categories did not reflect scale differences as much as older children's.

Moreover, we found a general developmental progression of initially broad and overlapping fruit categories to more differentiated distributions. These results are intriguing, as they are at odds with previous work highlighting that younger children have less permissive category representations and tend to prefer more extreme category members as representative instances (see, for instance, Foster-Hanson & Rhodes, 2019). In this account, our results might reflect that younger children were noisier, producing more diffuse distributions. However, our work is consistent with the idea that younger children exhibit higher cognitive flexibility (Gopnik et al., 2017). Future research should assess these competing explanations, for example, by evaluating the $MCMC_p$ distributions we obtained in convergent tasks, such as feature generalization tasks.

These types of experiments could facilitate a deeper understanding of how children's categories develop and how category structure determines children's and adults' generalization and inference capabilities, showcasing the strengths of $MCMC_p$ as a developmental paradigm. More broadly, we see much potential in using $MCMC_p$ and linked variants in experiments beyond developmental studies, given its effectiveness in estimating population-level categories.

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada [funding reference number 2016-05552] and a Seed grant from Brown University [reference GR300294]. We thank Katherine McGinn, Liana Haigis, Yemi Hailemariam, and Naveen Abraham for their help in conducting the experiments. We thank the three anonymous reviewers for their insightful comments and suggestions.

References

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 434–455.

Foster-Hanson, E., & Rhodes, M. (2019). Is the most representative skunk the average or the stinkiest? Developmental changes in representations of biological categories. *Cognitive Psychology*, *110*, 1–15.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*, 381–386.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*(1), 256.

León-Villagrá, P., Ehrlich, I., Lucas, C., & Buchsbaum, D. (2022). Uncovering children's concepts and conceptual change. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

León-Villagrá, P., Klar, V. S., Sanborn, A. N., & Lucas, C. G. (2019). Exploring the representation of linear functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 41).

León-Villagrá, P., Otsubo, K., Lucas, C., & Buchsbaum, D. (2020). Uncovering Category Representations with Linked MCMC with People. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 42).

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories. *Cognitive Science*, *36*(1), 150–162.

McDuff, D. (2010). A human-Markov chain Monte Carlo method for investigating facial expression categorization. In *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 151–156).

Ramlee, F., Sanborn, A. N., & Tang, N. K. (2017). What sways people's judgment of sleep quality? A quantitative choice-making study with good and poor sleepers. *Sleep*, *40*(7).

Roberts, G., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*(2), 63–106.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.