Uncovering Category Representations with Linked MCMC with people

Pablo León-Villagrá[1,2] Kay Otsubo[1] Christopher G. Lucas[2] Daphna Buchsbaum[1]
[1]Department of Psychology, University of Toronto, Canada
[2]School of Informatics, University of Edinburgh, United Kingdom

Abstract

Cognitive science is often concerned with questions about our representations of concepts and the underlying psychological spaces in which these concepts are embedded. One method to reveal concepts and conceptual spaces experimentally is Markov chain Monte Carlo with people (MCMCP), where participants produce samples from their implicit categories. While MCMCP has allowed for the experimental study of psychological representations of complex categories, experiments are typically long and repetitive. Here, we contrasted the classical MCMCP design with a linked variant, in which each participant completed just a short run of MCMCP trials, which were then combined to produce a single sample set. We found that linking produced results that were nearly indistinguishable from classical MCMCP, and often converged to the desired distribution faster. Our results support linking as an approach for performing MCMCP experiments within broader populations, such as in developmental settings where large numbers of trials per participant are impractical.

Keywords: Experimental design, Markov chain Monte Carlo, categorization, representation

Introduction

A fundamental question in cognitive science is how categories are represented and how these representations are used to generalize to novel stimuli. Imagine seeing a persimmon for the first time and trying to infer what kind of fruit it is. Based on its size and color, one might think that it is a kind of orange. However, its shape and leaves might remind one of a tomato. To make an informed guess about what kind of fruit the persimmon is we need to weigh these hypotheses: how likely is it that an orange could have this type of shape and leaves, could a tomato have this color and texture, or is it a novel fruit altogether?

At the computational level, one can conceive of this categorization task as probability density estimation (Ashby & Alfonso-Reese, 1995; Jäkel, Schölkopf, & Wichmann, 2008). When deciding what kind of fruit the persimmon is, one must consider how likely it is to belong to known categories of fruits or a new fruit category. These categories can be thought of as mixtures of probability distributions over fruit features, where previously encountered stimuli determine the number and distributional properties of these categories (Rosseel, 2002; Sanborn, Griffiths, & Navarro, 2010). Thus, to better understand human categorization, it is crucial to obtain a fine-grained picture of how people represent these category distributions. One method to obtain these distributions directly is Markov chain Monte Carlo with people (MCMCP; Sanborn, Griffiths, & Shiffrin, 2010).

MCMCP takes inspiration from a prominent statistical method, Markov chain Monte Carlo (MCMC; for a general introduction, see MacKay, 2003). In MCMC, samples from an arbitrary distribution can be obtained by an iterative procedure. In each step, the current state of the MCMC sampler is compared against a proposed update. If the proposed update is accepted, the proposal becomes the new state of the sampler. For appropriate proposal and acceptance procedures, this method produces a chain of samples approximating the desired probability distribution.

As in MCMC, in MCMCP, participants are presented with a series of forced-choice questions between the current state and a proposed state. Each option is a proposed example of the category of interest. Participants are asked to select the more likely category member, and given their choice, the state of the sampler is updated. Sanborn, Griffiths, and Shiffrin (2010) showed that these choices correspond to a statistically valid acceptance procedure, so, after enough iterations, samples from MCMCP correspond to samples from the participants' category representations (the target distribution).

MCMCP offers several advantages over alternative methods for eliciting psychological spaces, such as multidimensional scaling (MDS; Torgerson, 1965; Shepard, 1980). For example, in MCMCP, the experimenter does not need to specify all experimental stimuli *a priori*, which makes it possible to explore categories with complex structures and more than one or two relevant features (Martin, Griffiths, & Sanborn, 2012). Furthermore, by relying on forced choices, MCMCP does not require participants to understand and express category structure or stimulus similarity explicitly. This makes the method a potentially interesting tool for the study of non-verbal groups like young children or non-human animals.

However, MCMCP typically requires hundreds or thousands of samples to capture the structure of a category (Sanborn, Griffiths, & Shiffrin, 2010; McDuff, 2010), so participants must perform repetitive judgments over long sessions, especially for complex or high-dimensional categories.

Previous MCMCP experiments have adapted strategies from MCMC methods to reduce experimental duration, for example, creating a more efficient sampling space (Hsu, Martin, Sanborn, & Griffiths, 2019), adapting more efficient sampling schemes to experimental paradigms (Blundell, Sanborn, & Griffiths, 2012), or using specialized proposal schemes (Sanborn, Griffiths, & Shiffrin, 2010; Ramlee, Sanborn, & Tang, 2017; León-Villagrá, Klar, Sanborn, & Lucas, 2019).

Another way to improve MCMCP experiments is to link several participants to produce one shared distribution (Martin et al., 2012; Ramlee et al., 2017). Here, rather than each participant generating a full set of samples, participants complete a shorter number of trials, using the final samples of a previous participant as their initial state. This setup allows multiple participants to provide a single sample set, reducing

the number of trials needed per participant.

While it has been previously noted that this procedure might increase power at the expense of obscuring individual differences (Ramlee et al., 2017), no systematic analysis of the trade-off introduced by linking participants has been performed. As a result, MCMCP remains infeasible as an experimental method for populations for which the long and repetitive choices are too taxing.

We conducted the first direct comparison of both procedures. In Experiment 1, we contrasted the quality of posterior distributions for the same categorization task, generated from both linked and unlinked MCMCP. In Experiment 2, we asked a separate group of participants to rate samples obtained from both conditions. This allowed us to establish that both methods produced good examples of fruits from each category, and to determine whether there were meaningful differences between procedures.

Experiment 1: Linked vs. Unlinked MCMCP

In Experiment 1, we compared posterior distributions obtained via both methods for three fruit categories: apple, orange, and grape. Our unlinked condition was a replication of Sanborn, Griffiths, and Shiffrin (2010), in which a small number of participants each made a large number of choices, independently of one another. In our linked condition, each participant made a smaller number of choices contributing to one of several sample sets. We hypothesized that both methods would produce similar distributions and psychologically similar exemplars in all fruit categories, which would validate the use of linking in MCMCP experiments.

Participants

Participants were 131 non-colorblind adults, from within the city of Toronto (8 in the unlinked condition, 123 in the linked condition; M_{age} = 25.10, SD_{age} = 9.48, 91 female, 38 male, 2 other). An additional 10 participants were excluded, either due to technical issues (n = 6), or low acceptance rates[1] (n = 4). In the linked condition, to match the unlinked condition's data, participants were collected until eight sets of trials were completed. On average, 15 participants were needed to complete one set in the linked condition.

Participants in the unlinked condition were compensated $10 per hour ($M$ = 3 hours; range: 2.28-3.40 hours) and were allowed to complete the study over multiple two-hour sessions. Participants in the linked condition were compensated $5 for 15 minutes of participation and had to have an acceptance rate lower than 42% after participating – two standard deviations above the average acceptance rate after 15 minutes in the unlinked condition.

Materials

The experiment was presented on a 13-inch Macbook Air laptop. The stimuli were stylized images of fruit, as in Sanborn, Griffiths, and Shiffrin (2010). The fruits were generated by calculating the convex hull over a set of three circles. Varying the radii of these circles, as well as the horizontal and vertical distance between them, created a set of complex shapes that resembled fruits (see Figure 1). Finally, three parameters determined the color of the fruit (hue, saturation, and lightness).

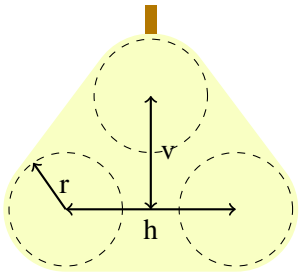

Figure 1: Stimuli were fruits that could be programmatically generated by specifying radii (r), vertical (v) and horizontal (h) length, and three color parameters.

Parameter ranges were taken from Sanborn, Griffiths, and Shiffrin with two exceptions: a 1.5 increase of radius and a 0.9 decrease in lightness. We increased the radius range, as additional analysis of data shared by Sanborn, Griffiths, and Shiffrin suggested that participants preferred slightly rounder fruits than those presented in their study. We decreased lightness to allow better visibility of the fruit on a white plate. Each fruit was topped with a brown stem to indicate the fruit's orientation to participants.

Procedure

On each trial, two fruits were displayed on top of white plates equidistant from the center of a black screen with instructions stating "Pick the [*fruit*]" for one of the three fruit categories: apple, orange, or grape, see Figure 2.

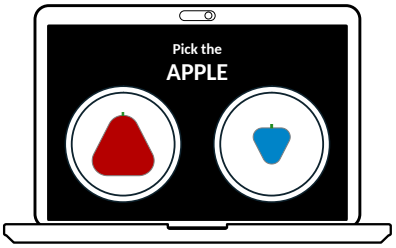

Figure 2: On each trial, participants chose which of the two options most resembled one of the three fruit categories.

For every trial, one fruit was the current state of the MCMC chain, while the other fruit was a sample from the proposal distribution (described below). The side on which each fruit appeared was randomized.

The proposal distribution was a mixture of a Gaussian distribution centered on the six current parameter values and two uniform distributions for shape and color parameters, $w = (0.8, 0.1, 0.1)$. The standard deviation of the Gaussian

[1] The acceptance rate is the proportion of trials in which the participant chooses the proposed fruit over the current.

was set to cover 0.07 of each parameter range. The uniform distributions allowed for the proposals to make large jumps in the parameter range, to avoid being stuck in isolated peaks. The uniform distributions sampled across either the three color or the three shape parameters while keeping the other parameters fixed. Any sample outside the valid parameter range was automatically rejected without being shown to participants; the current state was recorded, and the proposal counted as a rejection. A new proposal was then generated until all parameter values were within range.

Following Sanborn, Griffiths, and Shiffrin (2010), each fruit category had three independent chains, and participants completed 667 trials per chain, making a total of nine Markov chains and 6003 trials in a single sample set. Throughout the experiment, trials from the nine chains were presented in an interleaved sequence. The initial states for each chain were taken from Sanborn, Griffiths, and Shiffrin. All three fruits started from the same three initial states. Participants in the unlinked condition each completed an entire sample set, while participants in the linked condition completed as many trials as possible within a 15-minute session (M_{trials} = 393).

To implement linking, the next participant in the set continued from wherever the previous participant left off. In other words, a participant's last response on each of the nine chains became the first for the next participant in that set. Once the 6003 trials were completed for a set, the session automatically ended, and the next participant began a new set.

Results

Before analyzing the posterior distributions, it is critical to evaluate if the sampling process reliably captured participants' representation of fruit categories. Several diagnostics have been proposed in the statistical and MCMCP literature for whether an MCMC chain has accurately approximated its target distribution (Gelman et al., 2013; Sanborn, Griffiths, & Shiffrin, 2010). Here, we focus on evaluating two crucial characteristics of a reliable approximation: the number of (uncorrelated) samples obtained from the target distribution and the degree to which participants explored the fruit categories.

Usually, MCMC chains start at random locations – here, random fruit parameters that may not specify anything resembling a fruit – and the sampler has to move through the sample space until it lands in the area where the target distribution is concentrated. Then subsequent samples tend to stay in that region, and the chain is said to have converged. Thus, it is common practice to remove the samples before the chain has converged, often referred to as the burn-in period. We determined the length (in samples) of burn-in for each fruit category individually for each sample set, by incrementally calculating multivariate potential scale reduction factors (MPSRF)[2], a common metric of convergence from the statistics literature (Brooks & Gelman, 1998). Low MPSRF values indicate that a sequence of samples has converged. We used the point with the lowest MPSRF factor to determine the unique burn-in point for each set's category (for details, see Sanborn, Griffiths, and Shiffrin (2010)). We report samples after burn-in for all posterior distributions.

We observed no significant difference between the average length of burn-in between conditions, (M_{unlinked} = 158, SD_{unlinked} = 146; M_{linked} = 109, SD_{linked} = 110; t = 1.32, p = .01)[3]. On the other hand, average MPSRF values were significantly lower in the linked condition (M = 1.34, SD = 0.18) than in the unlinked condition (M = 1.89, SD = 0.93, t = 2.85, $p < .001$) suggesting that convergence was achieved within fewer trials when linking participants.

In addition to ensuring that the MCMCP chains converged, we also needed to verify that participants produced a sufficiently large number of independent, uncorrelated samples in order to obtain reliable approximations of the posterior distributions. However, since every state in a chain depends on the previous one, MCMC samples are correlated. A common way of estimating the number of independent samples in an MCMC chain is to estimate the effective sample size (ESS; Gelman et al., 2013). We calculated per-parameter ESS values for each of the nine chains (three per fruit category) in each sample set, and compared ESS values across conditions. ESS did not differ significantly between the unlinked and linked conditions, (M_{unlinked} = 10.66, SE_{unlinked} = 0.5, M_{linked} = 11.52, SE_{linked} = 0.67, t = -1.02, p = .16).

Given that samples with out-of-range parameters were automatically rejected, the number of samples was often higher than the total number of trials seen by participants. We, therefore, discuss acceptance rates in two ways. First, including these automatic rejections, as they are diagnostic for the sampling process (total acceptance rates), and second, excluding automatic rejections (human acceptance rates), as this corresponds to the proportion of proposals the participants accepted and thus is diagnostic for the psychological validity of the method.

For all convergence diagnostics per fruit category, see Table 1. Total acceptance rates (calculated per fruit) were relatively low in both conditions, compared to the recommended 20 - 40% (Roberts, Gelman, & Gilks, 1997). However, our rates were similar to those reported in Sanborn, Griffiths, and Shiffrin (2010). Human acceptance rates were closer to the recommended range. These results suggest that our proposal schemes were successful in that they allowed the participants to explore the category distribution efficiently.

Posterior Distributions

We found that the linked experimental design produced faster convergence and comparable numbers of samples as the unlinked condition. However, our main interest was in the cate-

[2]We considered the MPSRF for the full length and iteratively calculated MPSRF for chain lengths up to half of the total length of the shortest chain.

[3]We report bootstrapped, two-sample t-tests (Efron & Tibshirani, 1994). The p value corresponds to the proportion of permutations at least as extreme as the observed t. For all tests the number of permutation was set to 10,000. Traditional unequal variance, two-sample t-tests resulted in virtually identical results.

Table 1: Total number of samples (including automatic rejections) and effective sample sizes. The fruit category ESS was obtained by first calculating ESS for each set and then summing over all sample sets. We report total acceptance rates (*act*) and human acceptance rates (*ach*).

Condition	Fruit	N	M_{ESS}	M_{act}	SE_{act}	M_{ach}	SE_{ach}
Unlinked	Apple	22552	185	9%	10%	14%	12%
	Orange	18104	211	10%	11%	14%	12%
	Grape	17727	179	15%	13%	21%	14%
Linked	Apple	19787	166	13%	12%	19%	14%
	Orange	19729	249	10%	11%	14%	12%
	Grape	19317	206	17%	13%	23%	15%

gory structures that both methods uncovered.

To visualize the distribution over fruit categories for both conditions, we jointly embedded the samples in a two-dimensional plane using PCA, a common dimensionality reduction technique (see Figure 3). Both conditions produced qualitatively similar spaces. In both conditions, the horizontal dimension appears to broadly separate grapes relative to apples and oranges. In contrast, the vertical dimension has less overall separation but captures a relatively distinct clustering of oranges relative to the other two fruits.

The sample distributions in the linked and unlinked conditions qualitatively agreed in terms of their overall shapes and medians for the individual parameters (see Figure 4). Furthermore, our distributions closely matched those obtained by Sanborn, Griffiths, and Shiffrin (2010), suggesting that both linked and unlinked conditions produced comparable fruit distributions and reproduced their results.

While the overall distribution for apples was practically identical, some minor differences were apparent for oranges and grapes. First, the posterior density for oranges was more concentrated on larger radii in the linked condition, indicating that participants selected slightly larger oranges. More interestingly, while the median for the vertical parameter for grapes was similar across conditions, both distributions differed considerably in their overall shape. While the posterior distribution in the linked condition was unimodal and centered on zero – producing rounder grapes – the unlinked condition exhibited multiple modes. As a result, the posterior distribution exhibited higher density at the edges of the parameter range, corresponding to more oblong grapes.

Discussion

We compared the posterior distributions obtained by two different MCMCP methods: a classical MCMCP experiment, and a linked experiment. The linked experiment produced faster convergence, comparable sample sizes, and, qualitatively, nearly indistinguishable category distributions. However, we did not show that these category distributions were also psychologically equivalent. For example, the small differences in grape shapes and orange sizes that we found could amount to large perceptual differences. In Experiment 2, we ran a preregistered follow-up to assess the psychological representativeness of our samples directly.

Experiment 2: Subjective Fruit Ratings

To establish that both linked and unlinked methods generated equally-representative fruits, we asked participants to rate fruit samples taken from both conditions on how much they resembled each of the three fruit categories. We predicted that fruits from both experimental conditions would produce similar ratings, and fruits taken from a specific category would be rated as better examples of that category than the two alternatives (e.g., true apples would be rated as the best examples of apples, and as better examples of apples than of oranges or grapes).

Participants

A power analysis based on a separate set of 10 pilot participants established that at least 40 participants were required to achieve a power of $\geq 80\%$ with an α level of .05 to detect main effects and interactions. We recruited 40 non-colorblind participants from within the city of Toronto (M_{age} = 22.5, SD_{age} = 4.93, 27 female, 12 male, 1 other). Participants were paid \$5 to complete the task (M = 22.15 minutes, SD = 4.79, min = 13.29, max = 34.12). All participants passed our preregistered exclusion criteria[4].

Materials

We created four non-overlapping sets of 192 fruits by sampling from the fruits produced in Experiment 1 (after burn-in). In each set, we sampled 96 fruits from each condition (linked and unlinked), equally split over fruit categories (32 from each) and sample sets (four from each). As in Experiment 1, fruits were pictured centered on a plate. Fruits were presented one at a time, and three 7-point scales were positioned below, each with the corresponding rating prompt (e.g. "How much does this fruit resemble an apple?"). The order of the three rating scales was randomized across participants. Participants were randomly presented one of the four sets of fruits.

Procedure

Participants were instructed to rate each fruit in terms of how much it resembled each of the three fruit categories. To proceed to the next fruit, participants had to select a value for each of the three rating prompts (0-not at all, 6-exactly). The 192 fruits were presented in random order. Once the participants had completed the 192 ratings, they completed a short demographics survey.

Results

Fruits were consistently rated higher in terms of their true category than the two alternatives. Interestingly, oranges were rated relatively highly as apples, but apples were not rated

[4]Participants had to answer a comprehension check correctly and were excluded if they selected the same value on all three rating scales on more than 90% of the total trials.

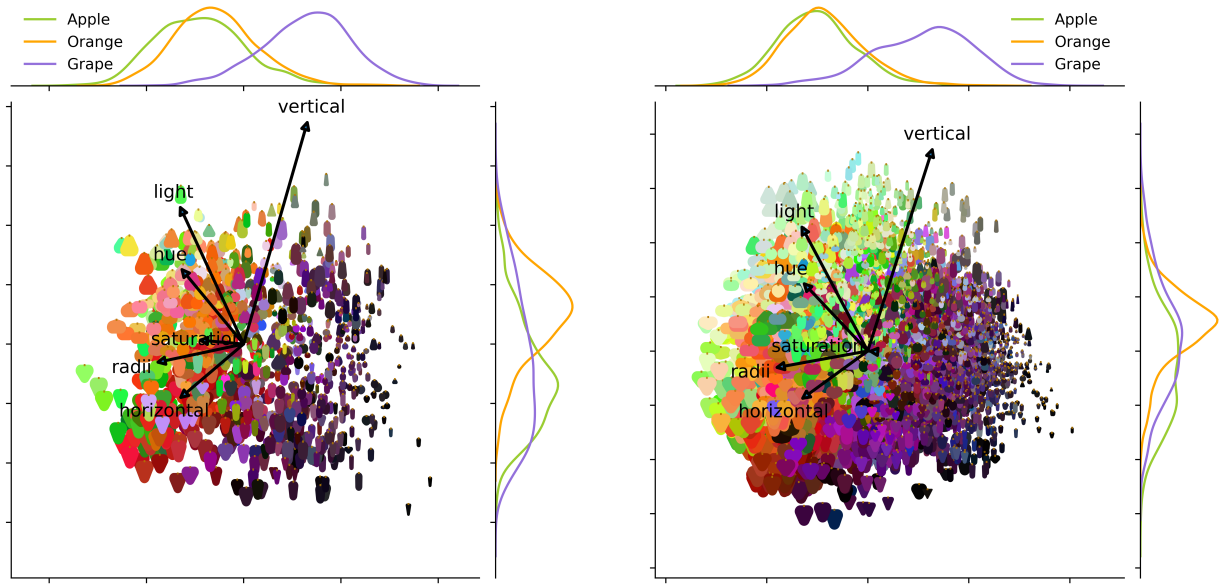

Figure 3: Samples obtained for the unlinked (left) and linked conditions (right). We calculated a shared two-dimensional representation of the data via its principal components.

highly as oranges. For a direct comparison of average fruit ratings, see Figure 5.

To examine how ratings differed across fruit categories and experiments, we fitted a $2 \times 3 \times 3$ Bayesian generalized linear model[5], experiment condition (linked vs unlinked) $\times$ true fruit category (apple, orange, grape) $\times$ rating category (apple, orange, grape), to participants' ratings. To account for variation in overall rating standards, the models had random per-participant intercepts. Since participants potentially varied in their rating standards for individual fruit categories, we also specified random slopes across fruit categories.

To test whether both experiments produced equally good fruits, we compared models with and without the effect of experimental condition. To test whether fruits were better examples of the true category than the alternatives, we compared models with and without an interaction for rating and true fruit.

This resulted in a total of five models: the full $2 \times 3 \times 3$ model, a model without the effect of experiment condition ($fruit \times rating$), a model without the interaction of rating and true fruit ($exp \times (fruit + rating)$), a model with only $fruit + rating$, and the null-model ($y \sim 1$). For all models, we specified weakly informative priors and ran two chains of 40,000 MCMC iterations.

All models accounted for the data better than the null-model (all BF $>$ 300). Contrasting the full model and $fruit \times rating$ suggested that the model without a factor for experimental condition accounted better for the data (BF $>$ 300). Finally, including the interaction $fruit \times rating$ improved model likelihoods (both BFs $>$ 300). Thus, we found strong evidence for the absence of an effect of experimental condition on fruit ratings, and an interaction of $fruit \times rating$.

Contrast analysis confirmed that the true fruits were rated more highly as their category than either alternative fruit. True apples were rated as more apple-like than either oranges (Mdn = 0.82, CI^{95} = [0.74, 0.9]) or grapes (Mdn = 1.22, CI^{95} = [1.15, 1.3]). Similarly, true oranges were rated higher as oranges than apples (Mdn = 0.89, CI^{95} = [0.82, 0.97]) or grapes (Mdn = 1.06, CI^{95} = [0.98, 1.13]). Finally, grapes were also scored higher as grapes than apples (Mdn = 1.01, CI^{95} = [0.93, 1.09]) or oranges (Mdn = 1.12, CI^{95} = [1.05, 1.20]). All contrasts can be considered highly significant, since no CI^{95} estimate overlapped with the 90% region of practical equivalence (ROPE; Kruschke, 2018).

To examine the small differences between experimental conditions, we performed a contrast analysis for the full model. We found that oranges in the linked condition were rated slightly higher as oranges than in the unlinked condition (Mdn = 0.23, CI^{95} = [0.13, 0.34], $ROPE_{90\%}$ = 0%). All other estimated effects did not suggest a significant difference in ratings across conditions (all $ROPE_{90\%} > 38\%$).

Discussion

We found that the samples in both linked and unlinked conditions were highly rated in terms of their corresponding fruit categories. This suggests that both designs produced equally good category distributions. Our results also revealed interesting asymmetries in the psychological fruit space. For instance, oranges were rated relatively highly as apples, but ap-

[5]Our preregistration specified analyzing the results with a linear mixed model. At the suggestion of the reviewers, we performed a Bayesian analysis to find evidence for the *absence* of an effect of experimental condition. The results of the linear mixed model were practically identical to the results presented here.

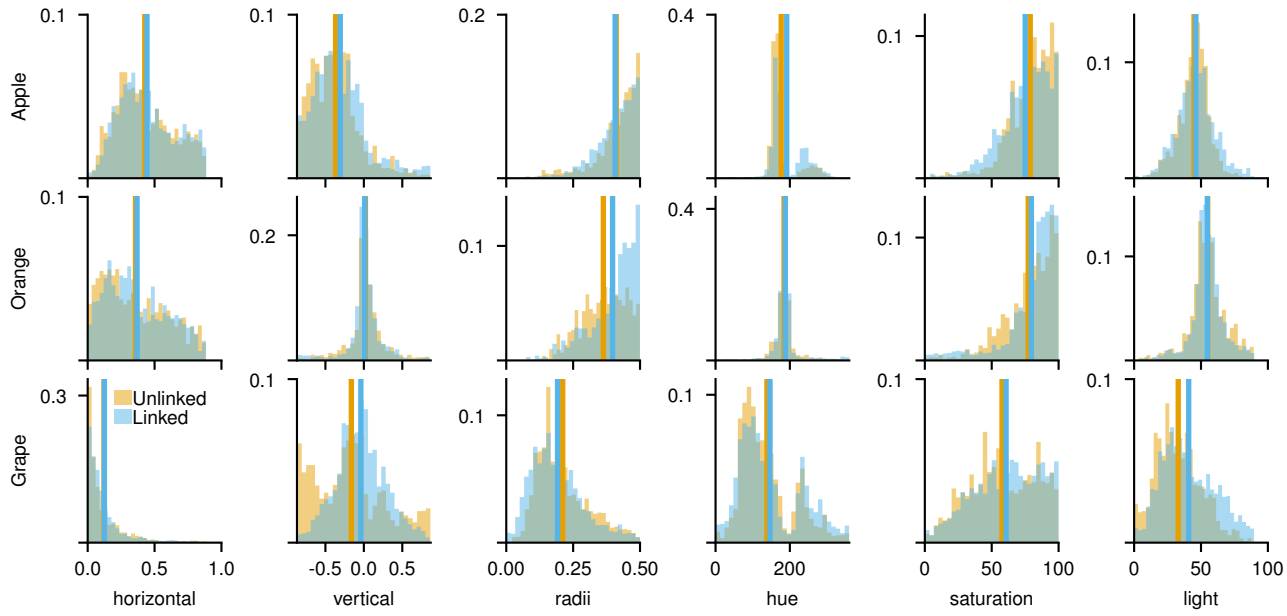

Figure 4: Histograms for linked and unlinked conditions, as well as medians (vertical lines). In general, both conditions resulted in very similar parameter distributions, corresponding closely in terms of modes, overall shapes, and moments.

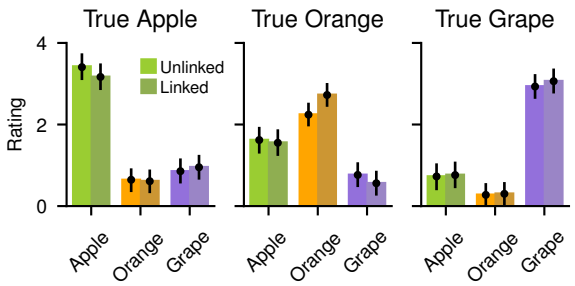

Figure 5: Average rating, estimated medians, and 95% credible intervals for each true fruit category across linked and unlinked conditions (linked shown in a darker shade).

ples were not rated highly as oranges, perhaps because they have somewhat similar shapes, but oranges are more specific in color. Similar to the results of our two-dimensional embedding, it seems that apples and oranges overlap in crucial parameters, as well as in their psychological perception. Altogether, these results corroborate our MCMCP manipulation – both linked and unlinked designs produced very similar distributional samples and psychologically representative category members.

Conclusion

We found that a linked MCMCP design converged faster, resulted in similar effective sample sizes, and did not differ considerably in terms of its overall distribution or psychological plausibility compared to a standard unlinked design.

Since linked MCMCP does not require long and repetitive experimental sessions, this suggests that it is possible to use MCMCP with experimental populations for which unlinked designs would be prohibitively taxing. Given its simplistic design and minimal linguistic and motor-control requirements, we see potential in adopting the method to explore mental representations of young children and even infants. Given that linked MCMCP offers the prospect of a single methodology suitable for a wide age range, this provides the opportunity to investigate developmental trajectories at an unprecedented level of detail.

However, adopting linked MCMCP experiments requires careful consideration of individual differences and overall category variability. While it is plausible that the fruit categories in our experiment were relatively homogeneous across adult populations, infants and children within the same age group can vary widely in their category knowledge (Smith, 2003; Bornstein & Arterberry, 2010; Slone, Smith, & Yu, 2019). Similarly, adult mental representations might differ given different levels of expertise (Chi, Feltovich, & Glaser, 1981; Medin, Lynch, Coley, & Atran, 1997; Bailenson, Shum, Atran, Medin, & Coley, 2002).

Given large variability in the linked chains, MCMCP would wash out individual differences and, similarly to other category learning tasks, only capture group-level representations. However, more informed linking strategies might alleviate these issues. For instance, vocabulary growth, particularly for object-names, can be a better predictor of children's category knowledge than age (Gopnik & Meltzoff, 1992; Arterberry & Bornstein, 2002; Smith, 2003). Future work should thus investigate ways of using these markers to link MCMCP chains optimally.

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada [funding reference number 2016-05552]. We thank Adam Sanborn for sharing the reference data set and for help throughout this project. We also thank the four anonymous reviewers for their thoughtful comments and suggestions. Finally, we thank Isaac Ehrlich, Joon Park, and Ben Prystawski for their help running these experiments.

References

Arterberry, M. E., & Bornstein, M. H. (2002). Variability and its sources in infant categorization. *Infant Behavior and Development*, *25*(4), 515–528.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition*, *84*(1), 1–53.

Blundell, C., Sanborn, A., & Griffiths, T. (2012). Look-ahead Monte Carlo with people. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Bornstein, M. H., & Arterberry, M. E. (2010). The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. *Developmental Psychology*, *46*(2), 350–365.

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

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

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

Gopnik, A., & Meltzoff, A. (1992). Categorization and Naming: Basic-Level Sorting in Eighteen-Month-Olds and Its Relation to Language. *Child Development*, *63*(5), 1091–1103.

Hsu, A. S., Martin, J. B., Sanborn, A. N., & Griffiths, T. L. (2019). Identifying category representations for complex stimuli using discrete Markov chain Monte Carlo with people. *Behavior Research Methods*, 1–11.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.

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.*

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).

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, *32*(1), 49–96.

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.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.

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.

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science*, *22*(6).

Smith, L. B. (2003). Learning to Recognize Objects. *Psychological Science*, *14*(3), 244–250.

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, *30*(4), 379–393.