Supplementary Materials for

Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians

Stephen Ferrigno*, Samuel J. Cheyette, Steven T. Piantadosi, Jessica F. Cantlon

*Corresponding author. Email: sferrigno@g.harvard.edu

Published 26 June 2020, Sci. Adv. 6, eaaz1002 (2020)
DOI: 10.1126/sciadv.aaz1002

The PDF file includes:

- Supplementary Materials and Methods
- Supplementary Results
- Figs. S1 to S9
- Legend for movie S1
- References

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/26/eaaz1002/DC1)

- Movie S1
Supplementary Materials
Materials and Methods

Bayesian model strategies

We formally defined a strategy as a sequence of task-relevant operations composed of three primitives: O, C, and M. O and C choose a random open and closed bracket from the screen; a parameter ($\gamma$) determines how biased each one of these is towards a specific open or closed bracket. M searches through memory for the most recent unmatched bracket and then returns the opposing bracket of the same type. For example, the strategy $OOMM$ first chooses an open at random, then another open, then matches the second open, then matches the first open–this strategy correctly outputs only center-embedded recursive sequences. The strategy $OOCC$, on the other hand, is equally likely to generate center-embedded and crossed responses, $([], []), ([[]])$ and $([]), since C chooses an available “close” at random, regardless of whether it matches the most recent open. An $OOCC$ strategy is not hierarchical or embedded. This $OOCC$ strategy is simply knowing that the open brackets should come earlier than the close brackets, and would yield the same response pattern as an ordinal strategy. Similarly, $OMOM$ would generate sequences like $([])$, consistent with an associative chain strategy. We define a recursive strategy as one that results in choosing two open brackets and their matching types in order (e.g., $OOMM$).

The strategies allow for biases towards one specific bracket or another, as well as variable levels of noisiness. For instance, $OOMM$ could generate center-embedded structures that are biased to begin with “[” rather than a random open bracket; or it could make mistaken bracket-choices frequently. The analysis includes a parameter for each participant's “noisiness” as the probability that they make a choice inconsistent with the strategy they meant to use on a given trial. With this parameter, some incorrect responses, such as “([[])”, the participants have actually intended a correct center-embedded response, such as “([[])”. The converse is true as well: some center-embedded responses may have been an accident. By jointly fitting the participants’ intentions, noisiness, and bias together, the model can provide a more complete account of the full range of responses allowing us to more precisely determine what participants have actually learned.
Not including such biases and noise, the hypothesis space over which inference was performed consisted of all unique strategies that output 4 brackets, of which there are 12. Finally, we considered it likely that many participants used mixtures of strategies. To handle this, the model infers a distribution of response strategies for each subject.

The model was constructed to respect the grouping in the data—namely that individuals provide multiple responses and that multiple individuals come from each group (monkeys, US kids, US adults, and Tsimane’ adults). This allows us to determine what unique biases individuals of each group may have towards certain strategies. This analysis required three group-level latent parameters and three individual latent parameters which were partially pooled within groups. Figure S1 shows the structure of the model with each parameter in plate-diagram format. The three group variables inferred were: \( \beta_g \), a mean distribution over strategies; \( \alpha_g \), a clustering parameter specifying the homogeneity of the population around the mean distribution; and \( \eta_g \), a noise parameter specifying how often mistakes were made in following a strategy. The three variables inferred for participants were: \( \theta_p \), a distribution over strategies, dependent on \( \alpha_g \) and \( \beta_g \); \( \nu_p \), a noise parameter dependent on \( \eta_g \); and 3) \( \gamma_p \), a term capturing bias in choosing one type of bracket over another. The responses \( R_p \) of each individual, represented by counts of bracket-choices, are then drawn for each participant. This model’s structure, treating individuals as mixtures of strategies, is similar to Latent Dirichlet Allocation (43). We trained this model with the gradient-based MCMC algorithm NUTS (44). Since \( \beta_g \) and \( \theta_p \) are probability vectors representing the posterior probability over each strategy by each group and participant, it is easy to extract the probability that they were each using a recursive strategy in particular.

**Bayesian Model Structure**

The Bayesian model was structured hierarchically, with participants partially pooled by their group (monkey, US kids, US adults, Tsimane’ adults). The group variables inferred were \( \beta_g \), \( \alpha_g \), and \( \eta_g \). \( \beta_g \) is a probability vector over strategies, representing the group mean likelihood of using each of twelve possible strategy, and is drawn from a Dirichlet with a uniform prior. \( \alpha_g \) is a clustering parameter specifying how sparse or tightly the participants in a group cluster around their \( \beta_g \). It is drawn from an
Exponential distribution with parameter 0.1, representing a bias towards homogeneity of groups. $\eta_g$ is a scalar specifying the group-mean noise of implementing strategies, drawn from a Beta distribution with parameters $\alpha = 1, \beta = 9$, specifying a prior towards low levels of noise. This is because it is best to explain differences in responses in terms of strategy selection and rely on noise to explain differences in the data only if it’s necessary (both for explanatory purposes and to prevent over-fitting).

The participant-level variables were $\gamma_p$, $\theta_p$, $\nu_p$, and $R_p$. $\gamma_p$ determines how biased each strategy was towards starting with a particular open bracket, and is drawn from a uniform Beta distribution ($\alpha = 1, \beta = 1$). We did not determine bias for closed-brackets as participants across groups almost always picked open-brackets first, and we were most interested in differentiating between center-embedded and crossed responses, both of which start with an open bracket. More specifically $\gamma_p$ the first choice of an open bracket, $\eta$ specifies how likely that bracket is to be of one particular kind, e.g. “[” rather than “(”. $\gamma_p$ is drawn from a Beta distribution with a uniform prior. $\theta_p$ is a distribution over strategies, drawn from a Dirichlet with prior $\alpha^T \beta_g$. Given a set of strategies $S$, for each participant $p$ in group $g$, the model in full is below:

$$
\beta_g \sim \text{Dirichlet}(1)
$$

$$
\alpha_g \sim \text{Exponential}(0.1)
$$

$$
\eta_g \sim \text{Beta}(1, 9)
$$

$$
\gamma_p \sim \text{Beta}(1, 1)
$$

$$
\theta_p \sim \text{Dirichlet}(\alpha_g^T \beta_g)
$$

$$
\nu_p \sim \text{Beta}(1 - \eta_g, \eta_g)
$$

$$
R_p \sim \text{Multinomial}(F(\theta_p^T S, \nu_p, \gamma_p))
$$

The function $F$, used to calculate $R_p$, adds noise and bias to the responses of strategies. Noise is added to each strategy’s responses by determining every possible response’s likelihood of having resulted from following that strategy. More specifically,
responses that are more similar (have more overlap) to those that are intentionally output by the strategy are more likely to have been generated by that strategy. We define the distance \( D \) between two responses as the total number of places in their output they diverge — e.g., \( D=1 \) for \([()\)] and \([(())\)] and \( D=2 \) for \([()\)] and \([(())\)]. The probability that one output was “supposed” to be another output but got corrupted given a distance \( D \) and a noise-level \( \eta \) is \( \eta^D \times (1-\eta)^{4-D} \). These probabilities are factored into each strategy by marginalizing over all the possible response pairs and their distances from the intended responses of a given strategy, re-weighting each based on the corresponding likelihood of corruption. Bias is added into each strategy by up-weighting one open-bracket over another by a factor of \( \gamma \). So if \( \gamma \) is 0.2, for example, the first time \( O \) is called, one open bracket is called with probability 0.8 and the other is called with probability 0.2.

**Bayesian model training**

The Bayesian model was trained using PyMC3 (45), with the default MCMC algorithm NUTS. It was run for 2,000 steps with 500 tuning steps, and a thin of 10. The low number of samples is due to NUTS being a gradient-based MCMC technique, and thus requires many fewer steps to converge than classic MCMC algorithms. We ran two chains to test convergence, which we confirmed using standard diagnostics. 95% credible intervals for parameters were determined by taking the smallest range containing 95% of samples.

**Supplemental Results**

**Bayesian model results**

The full group-level means over strategies, represented by the parameter \( \beta_g \), is shown in Fig. S6. The inferred probabilities that individual participants were using each strategy, represented by the parameter \( \theta_p \), is shown in Fig. S7.

Using a spearman regression, we found a significant correlation between US kids’ performance on the memory task and their inferred memory noise from the model (spearman-\( \rho = -0.36, p < 0.05 \)). A plot of this effect is shown in Fig. S5, with a loess fit
to the points (because the relationship between the variables is, by necessity, non-linear).

There were differences between the group-level clustering parameter $\alpha$. A higher $\alpha$ value corresponds to individuals in a group more tightly clustering around their group mean — so having more similar strategies. US adults had by far the highest $\alpha$ ($M = 14.3$, $CI = [5.6, 27.2]$), followed by Tsimane’ adults ($M=6.5$, $CI=[2.6,18.9]$), US kids ($M = 3.6$, $CI = [2.6, 4.9]$), and then monkeys in experiment 2 ($M = 3.3$, $CI = [1.8, 5.7]$), see Fig. S8. It is intuitive that adults had the highest $\alpha$ — and thus were most tightly clustered — considering every adult was inferred to be using the strategy $OOMM$ with very high probability.

**Bayesian analysis of monkeys’ performance on Experiments 1 and 3**

We ran the same Bayesian analysis on monkeys’ responses in the first and third experiments to understand the extent to which they were using a recursive strategy to respond in either. The analysis of their performance in Experiment 1 provides evidence that they were not using a recursive strategy frequently. As shown in Fig. S3, the monkeys seemed to be using the non-recursive strategy $OOCC$ — generating both center-embedded and crossed responses — at a much higher rate ($M=0.39; CI=[0.20,0.53]$) than they used the recursive strategy $OOMM$ ($M=0.07, CI=[0.02, 0.15]$). Moreover, no individual monkey was inferred to use the strategy $OOMM$ at a rate above 0.08. The analysis further revealed that two of the monkeys made very few errors, with mean error rates on trials of 0.03 ($CI=[0.01,0.07]$) and 0.07 ($CI=0.03, 0.13$). However, the other monkey was inferred to be nearly random, with an error rate of 0.86 ($CI=[0.70, 0.95]$). This implies that two monkeys understood the task but did not generalize to a recursive strategy and the other monkey either did not understand the task or was not trying.

While the analysis of Experiment 2 provides evidence that monkeys eventually learned the recursive strategy with the same brackets, we were also interested in whether the monkeys generalized the recursive strategy to novel brackets in Experiment 3. As shown in Fig. S7, the monkeys used the non-recursive strategy $OOCC$ most frequently ($M=0.39; CI=[0.30,0.47]$), similar to the rate in Experiment 1. However, they also used the recursive strategy $OOMM$ at a much higher rate than in Experiment 1 ($M=0.17$, $CI=[0.09,0.26]$), though not quite as high as in Experiment 2. The posterior probability of
using $\text{OOMM}$ was higher than the prior for all three monkeys ($M_s = 0.14, 0.15, 0.20$), evidence that the monkeys were all explicitly using a recursive strategy at least some of the time. Beyoncé, the monkey who used a recursive strategy most frequently, was also the most noisy (as in the other two experiments), making at least one error on each trial with probability $0.42$ ($CI=[0.33,0.53]$). The other two monkeys again had insignificant inferred error rates by comparison, with both less than $0.05$.

**Additional analyses of monkey results**

Because multiple experiments were used to test for center-embedding in monkeys, we also wanted to test if across all experiments, the monkeys produced more center-embedded responses than crossed responses. We found that when we combined all three experiments, the monkeys were more likely to produce center embedded responses than chance (Binomial (two-tailed): Overall: $231/679, p < .001$) Additionally, the monkeys were more likely to produced center-embedded responses than crossed responses across all three experiments (Binomial (two-tailed): Overall: $231/409, p < .01$; see Fig. S1).
**Fig. S1.** The proportion of each possible response on transfer trials for monkeys, children, US Adults, and Tsimane’ adults. The monkey results show the total responses across all three experiments. Error bars represent the standard error of the mean.
**Fig. S2.** A plate diagram representation of the Hierarchical Bayesian Model. The group-level variables inferred are $\beta_g$, $\alpha_g$, and $\eta_g$. $\beta_g$ represents the group mean likelihood of using each strategy; $\alpha_g$ specifies how tightly the participants in a group cluster around their $\beta_g$; and $\eta_g$ represents the group-mean noise in implementing strategies. The participant-level variables are $\gamma_p$, $\theta_p$, and $\nu_p$. $\gamma_p$ determines how biased each strategy is towards starting with a particular open bracket; $\theta_p$ represents a participant’s likelihood of using each strategy; and $\nu_p$ specifies participants’ level of noise in responding.
Fig. S3. This plot displays the mean probability of using each strategy for monkeys in Experiment 1, as inferred by the Bayesian model. In the model, the vector of probabilities of strategy-use is represented by $\beta_g$. Error bars represent 95% credible intervals. The mean of the prior for each strategy is given by the dashed line ($y=1/12$).
**Fig. S4.** This plot displays the mean probability of using each strategy for monkeys in Experiment 3, as inferred by the Bayesian model. In the model, the vector of probabilities of strategy-use is represented by $\beta_g$. Error bars represent 95% credible intervals. The mean of the prior for each strategy is given by the dashed line ($y=1/12$).
Fig. S5. This figure displays the probability of making an error (y-axis) inferred by the Bayesian model on the y-axis, against the kids’ performance on the memory task (x-axis). The blue line shows the best-fit curve from a loess regression fit on the data.
**Fig. S6.** This plot displays the mean probability of using each strategy for each population, as inferred by the Bayesian model. In the model, the vector of probabilities of strategy-use for each group is represented by $\beta_g$. Error bars represent 95% credible intervals. The mean of the prior for each strategy is given by the dashed line ($y=1/12$).
**Fig. S7.** This figure shows the inferred probability of using each strategy (faceted) for individuals (x-axis, jittered). In the model, the vector of probabilities of strategy-use for each individual is represented by $\theta_g$. Error bars represent 95% credible intervals. The mean of the prior for each strategy is represented by dashed line.
**Fig. S8.** This figure displays the inferred $\alpha$ values from the model. These values indicate how similar the strategies of individuals in each population are. A higher $\alpha$ value indicates that participants are more tightly clustered around the group mean. Error bars are 95% credible intervals.
Fig. S9. The proportion of center-embedded, crossed, and tail-embedded responses for each monkey on the novel transfer trials by trial block across all three experiments. Each block represents 20 trials. Error bars represent the standard error of the proportion.
**Movie S1.** Example video of the recursion training and testing trials. This video shows a monkey completing one trial of each of the training trials as well as one of the non-differentially reinforced probe trials.
REFERENCES AND NOTES

1. S. Pinker, R. Jackendoff, The faculty of language: What’s special about it? *Cognition* **95**, 201–236 (2005).

2. N. Chomsky, *Syntactic Structure* (Mouton). (1957).

3. M. D. Hauser, N. Chomsky, W. T. Fitch, The faculty of language: What is it, who has it, and how did it evolve? *Science* **298**, 1569–1579 (2002).

4. N. I. Badler, R. Bindiganavale, J. C. Bourne, M. S. Palmer, J. Shi, Parameterized action representation for virtual human agents. in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, E. Churchill, Eds. (MIT Press, Cambridge, MA, 2000), pp. 256–284.

5. R. Jackendoff, *Language, Consciousness, Culture: Essays on Mental Structure* (MIT Press, 2007).

6. F. Lerdahl, R. Jackendoff, An overview of hierarchical structure in music. *Music Percept.* **1**, 229–252 (1983).

7. S. T. Piantadosi, J. B. Tenenbaum, N. D. Goodman, Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition* **123**, 199–217 (2012).

8. D. Everett et al., Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Curr. Anthropol.* **46**, 621–646 (2005).

9. A. Nevins, D. Pesetsky, C. Rodrigues, Pirahã exceptionality: A reassessment. *Language* **85**, 355–404 (2009).

10. R. Futrell, L. Stearns, D. L. Everett, S. T. Piantadosi, E. Gibson, A corpus investigation of syntactic embedding in Pirahã. *PLOS ONE* **11**, e0145289 (2016).

11. S. Dehaene, F. Meyniel, C. Wacongne, L. Wang, C. Pallier, The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron* **88**, 2–19 (2015).

12. C. R. Gallistel, *The Organization of Learning* (MIT Press, Cambridge, MA, 1990).

13. E. M. Brannon, H. S. Terrace, Ordering of the numerosities 1 to 9 by monkeys. *Science* **282**, 746–749 (1998).

14. E. M. Brannon, The development of ordinal numerical knowledge in infancy. *Cognition* **83**, 223–240 (2002).

15. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
16. M. D. Hauser, E. L. Newport, R. N. Aslin, Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition* **78**, B53–B64 (2001).
17. L. A. Heimbauer, C. M. Conway, M. H. Christiansen, M. J. Beran, M. J. Owren, Visual artificial grammar learning by rhesus macaques (*Macaca mulatta*): Exploring the role of grammar complexity and sequence length. *Anim. Cogn.* **21**, 267–284 (2018).
18. G. F. Marcus, S. Vijayan, S. B. Rao, P. M. Vishton, Rule learning by seven-month-old infants. *Science* **283**, 77–80 (1999).
19. J. Saffran, M. Hauser, R. Seibel, J. Kapfhamer, F. Tsao, F. Cushman, Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition* **107**, 479–500 (2008).
20. L. Wang, L. Uhrig, B. Jarraya, S Dehaene, Representation of numerical and sequential patterns in macaque and human brains. *Curr. Biol.*, **25**, 1966–1974 (2015).
21. R. Sonnweber, A. Ravignani, W. T. Fitch, Non-adjacent visual dependency learning in chimpanzees. *Anim. Cogn.* **18**, 733–745 (2015).
22. W. T. Fitch, M. D. Hauser, Computational constraints on syntactic processing in a nonhuman primate. *Science* **303**, 377–380 (2004).
23. T. Q. Gentner, K. M. Fenn, D. Margoliash, H. C. Nusbaum, Recursive syntactic pattern learning by songbirds. *Nature* **440**, 1204–1207 (2006).
24. D. C. Penn, K. J. Holyoak, D. J. Povinelli, Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* **31**, 109–130 (2008).
25. W. T. Fitch, Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Phys. Life Rev.* **11**, 329–364 (2014).
26. P. Perruchet, A. Rey, Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychon. Bull. Rev.* **12**, 307–313 (2005).
27. M. H. de Vries, P. Monaghan, S. Knecht, P. Zwitserlood, Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition* **107**, 763–774 (2008).
28. M. C. Corballis, Recursion, language, and starlings. *Cognit. Sci.* **31**, 697–704 (2007).
29. K. Abe, D. Watanabe, Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nat. Neurosci.* **14**, 1067–1074 (2011).
30. A. Rey, P. Perruchet, J. Fagot, Centre-embedded structures are a by-product of associative learning and working memory constraints: Evidence from baboons (Papio Papio). *Cognition* **123**, 180–184 (2012).

31. F. H. Poletiek, H. Fitz, B. R. Bocanegra, What baboons can (not) tell us about natural language grammars. *Cognition* **151**, 108–112 (2016).

32. M. Winkler, J. L. Mueller, A. D. Friederici, C. Männel, Infant cognition includes the potentially human-unique ability to encode embedding. *Sci. Adv.* **4**, eaar8334 (2018).

33. X. Jiang, T. Long, W. Cao, J. Li, S. Dehaene, L. Wang, Production of supra-regular spatial sequences by macaque monkeys. *Curr. Biol.* **28**, 1851–1859.e4 (2018).

34. J. Henrich, S. J. Heine, A. Norenzayan, Most people are not WEIRD. *Nature* **466**, 29 (2010).

35. H. S. Terrace, B. McGonigle, Memory and representation of serial order by children, monkeys, and pigeons. *Curr. Dir. Psychol. Sci.*, **3**, 180–185 (1994).

36. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis* (CRC press, Boca Raton, FL, 2014).

37. A. Ravignani, G. Westphal-Fitch, U. Aust, M. M. Schlumpp, W. T. Fitch, More than one way to see it: Individual heuristics in avian visual computation. *Cognition* **143**, 13–24 (2015).

38. E. Gibson, Linguistic complexity: Locality of syntactic dependencies. *Cognition*, **68**, 1–76 (1998).

39. H. Ginsburg, A. J. Baroody, *TEMA-3: Test of Early Mathematics Ability* (Pro-ed.), (2003).

40. E. Carrow-Woolfolk, *Test for Auditory Comprehension of Language* (DLM Teaching Resources, Allen, TX, 1985).

41. E. Gibson, J. Jara-Ettinger, R. Levy, S. Piantadosi, The use of a computer display exaggerates the connection between education and approximate number ability in remote populations. *Open Mind* **1**, 159–168 (2017).

42. J. Bahlmann, R. I. Schubotz, A. D. Friederici, Hierarchical artificial grammar processing engages Broca's area. *Neuroimage*, **42**, 525–534 (2008).

43. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).

44. M. D. Hoffman, A. Gelman, The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).
45. J. Salvatier, T. V. Wiecki, C. Fonnesbeck, Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.*, 2, e55 (2016).