recursion_writeup_3

Sam Cheyette

July 9, 2018

Many responses cannot be classified as center-embedded, tail-recusive, or crossed. Indeed, over 25% of all responses — and close to half of the monkeys responses — do not fall cleanly into one of those categories (notice that the bars in Figures 2 and 3 do not add up to 1). The previous analyses ignored these uncategorized responses, primarily using the relative proportion of crossed and center-embedded responses as a guage of a participant's learning. While this is entirely valid for its purpose, it does not account for a large portion of the data, and therefore can neither fully explain the variability of the responses nor precisely determine between-participant and between-group differences. So, to better understand the origins of the entire set of responses, we implemented a model to capture the process by which the responses were generated. More specifically, we performed a Bayesian data analysis to jointly infer the strategies used by each participant in the task to make each responses, as well as their noisiness (e.g. mis-presses, memory error, etc. . .) in implementing those strategies (Gelman et al., 2012). By modeling the strategies that were used by each participant to respond, we we can more precisely describe what participants learned; and by delineating which choices were *intentional* and *unintentional*, we can determine how they were hindered by mistakes.

We formally defined a strategy as a sequence of task-relevant operations [1]. On a given trial, the operations in a strategy are called sequentially until that trial is complete. The three primitive operations we defined are O, C, and M. O and C choose a random open and closed bracket from the screen; and a variable (γ) determines how biased each one is towards choosing *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 ([]), ([)], [(]) and [()] since C chooses an available "close" at random, regardless of whether it matches the most recent open. 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 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. A participant's "noisiness" refers to the probability that they make a choice inconsistent with the strategy they meant to use on a given trial. This means that for some incorrect responses, such as ([]], the participants have actually intended a correct center-embedded response, such as ([]). Though note that 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 strategies that output 4 brackets. Duplicate strategies — those that gave identical responses — were also removed, leaving 12 total strategies. Finally, we considered it likely that many participants used *mixtures* of strategies to produce responses over the course of the task, which would give rise to distributions of responses more complex than that of a single strategy. We deployed a Bayesian inference method to infer what mixture of strategies individuals and populations used.

The model was constructed to respect the hierarchical 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 members 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 5A shows the structure of the model with

[1] We restricted the number of operations in a strategy to 4, since only responses of length 4 were allowed in the experiment.

Figure 5

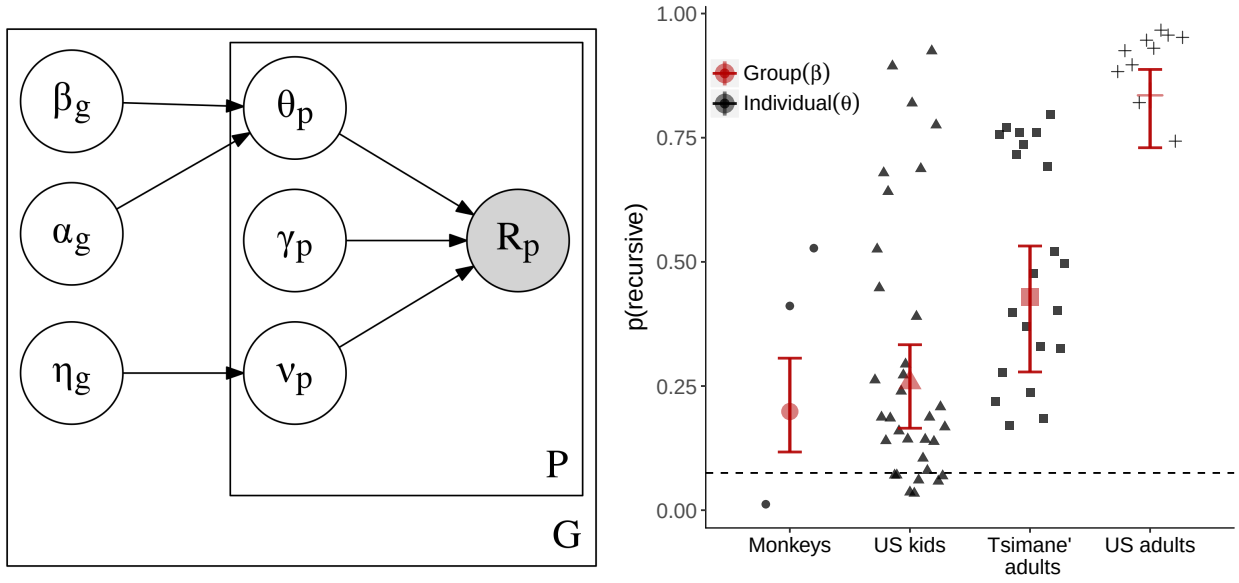


Figure 1: Panel 5A (left) displays a plate diagram representation of the Hierarchical Bayesian Model. The group-level variables inferred are β_g, α_g, and η_g. β_g represents the group mean likelihood of using each strategy; α_g specifies how tightly the participants in a group cluster around their β_g; and η_g represents the group-mean noise in implementing strategies. The participant-level variables are γ_p, θ_p, and ν_p. γ_p determines how biased each strategy is towards starting with a particular open bracket; θ_p represents a participant's likelihood of using each strategy; and ν_p specifies participants' level of noise in responding. Panel 5B (right) shows the probability of using a recurisve strategy for each group (red) and each individual in that group (black). Error bars around the group means represent the 95% credible interval.

each parameter in plate-diagram format. The three group variables inferred were: β_g, a mean distribution over strategies; α_g, a clustering parameter specifying the homogeneity of the population around the mean distribution; and η_g, a noise parameter specifying how often mistakes were made in following a strategy. The three variables inferred for participants were: θ_p, a distribution over strategies, dependent on α_g and β_g; ν_p, a noise parameter dependent on η_g; and 3) γ_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 (Blei, Ng, & Jordan, 2003). We trained this model with with the grdient-based MCMC algorithm NUTS (Hoffman & Gelman, 2011).

Since β_g and θ_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. Figure 5B shows the probability that individuals in each group were using a recursive strategy, both at the group-level (β_g) and for each individual in a group (θ_p). The prior probability of using a strategy on a given trial was 1/12 (≈ 0.08) for each of the 12 strategies. Each group was inferred to be more likely using a recursive strategy than would be a priori expected, as each group's recursive β_g was inferred to be very likely greater than 1/12 [2] The rank-order of the Maximum A Posteriori (MAP) values for β_g rank in order from US adults highest (M=0.83; CI=[0.72, 0.88]), followed by Tsimane' adults (M=0.42, CI=[0.28, 0.53]), US kids (M=0.25; CI=[0.17, 0.33]), and then monkeys (M=0.21; CI=[0.12, 0.33]). The individual MAP θ_p values, however, tell a more subtle story: the relatively low average recursive strategy use by monkeys (β_g) is heavily driven by a single monkey who had near-zero probability mass on the correct recursive strategy. This monkey was inferred to have used the strategy $OMOM$ approximately 71% of the time — generating "tail-recursive" responses instead. However, the monkey inferred to use a recursive strategy most often had a mean recursive θ_p value of 0.51 (CI=[0.28,0.66]), higher than 59% of human participants (76% of US kids and 62% of Tsimane' adults). The monkey with the next-most recursive strategy use is not far behind, with a θ_p of 0.41 (CI=[0.20,0.59]) — higher than 53% of human participants.

Memory constraints on recursive processing

The participant-level noise parameter ν_p in the Bayesian model specifies the probability that any given bracket-choice a participant made was unintended. The group-level noise parameter, η_g, specifies the probability of making a mistake aggregating over all participants in a group. We made the simplifying assumptions that 1) a mistake changes the intended bracket to one of the other three brackets at random; and 2) that each mistake is independent of other mistakes. This noise-model is not designed to account for every possible source of error separately, of which there are many (e.g. inattention, memory failure, mis-presses, etc...); rather, it was designed to be a general catch for responses that were unlikely to have been generated intentionally, without respect to their exact causes.

The Bayesian analysis revealed large differences between groups in the inferred amount of noise. Monkeys were inferred to have the highest levels of error, followed by US kids, Tsimane' adults, and then US adults. These differences are substantial: monkeys had an error rate of 0.075 on any given bracket choice, which corresponds to an error rate of 0.24 over all trials ($CI = [0.19, 0.28]$). This is over 80% higher than the error rate of US kids ($M = 0.16, CI = [0.12, 0.19]$, 140% higher than Tsimane' adults ($M = 0.10, CI = [0.07, 0.13]$), and 520% higher than US adults ($M = 0.04, CI = [0.02, 0.05]$). Figure 6A shows the probability that individuals in each group made an error on a given trial.

The differing levels of noise between groups can explain some of the difference in their ability to correctly and consistently center-embed. We compared the model's predictions with and without the noise parameters η_g and ν_p — holding the other inferred parameters constant — to determine the effect of noise on each group's performance. The results, displayed in Figure 6B., show that monkeys would center-embed with probability 0.41 (CI=[0.35, 0.48]) if they implemented their inferred strategies correctly, compared to their previous rate of 0.26. This is an increase in center-embedding rate of 57%, substantially higher than kids (12%), Tsimane' adults (7%) and US adults (7%). The absolute differences in center-embedding between monkeys

[2] In fact, that the recursive β_g is not 0 (or close to 0) can by itself be viewed as evidence that a recursive strategy was used. We used a comparison to the prior here to highlight that there is strong evidence specifically favoring a recursive strategy, as opposed to the case where no strategy has much evidential weight.)

Figure 6

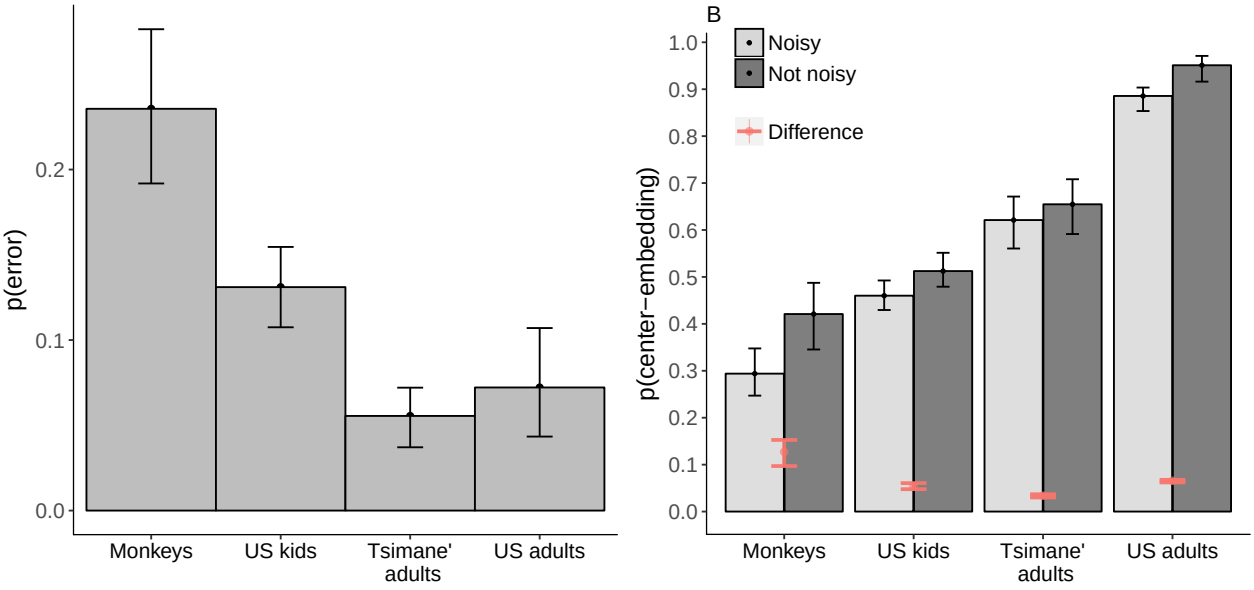


Figure 2: Panel 6A (left) shows the probability each group made an error implementing their strategy at least once in a trial, according to the results of the Bayesian analysis. Panel 6B shows the probability each group generates center-embedded responses, with noise included in the model (light gray bars, left) and excluded from it (dark gray bars, right). The red points represent the difference in center-embedding rates with and without noise for each group (i.e. the difference in the height of the bars). The error bars around the means in both panels represent the 95% credible intervals.

and the other groups would also diminish. For instance, the difference in rates of center-embedding between monkeys and US kids (removing US kids' errors as well) would drop 47% from 0.17 (CI=[0.14,0.18]) to 0.09 (CI=[0.06,0.13]).

Additionally, in US children we found that the inferred probability of making an error correlated with their memory performance ($\rho = -0.43; p = 0.02$; see Figure S4).

Supplementary Materials

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 β_g, α_g, and η_g. β_g is a probability vector over strategies, representing the group mean likelihood of using each strategy, and is drawn from a Dirichlet with a uniform prior. α_g is a clustering parameter specifying how sparse or tightly the participants in a group cluster around their β_g, and is drawn from an Exponential distribution with parameter 1. η_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 γ_p, θ_p, ν_p, and R_p. γ_p determines how biased each strategpy 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 well 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 γ_p the first choice of an open bracket, η specifies how likely that bracket is to be of one particular kind, e.g. "[" rather than "(". γ_p is drawn from a Beta

distribution with a uniform prior. θ_p is is a distribution over strategies, drawn from a Dirichlet with prior $\alpha_g^T \beta_g$.

Given a set of strategies S, for each participant p in group g, the model in full is below:

$$\beta_g \sim Dirichlet(1)$$
$$\alpha_g \sim Exponential(1)$$
$$\eta_g \sim Beta(1, 9)$$

\

$$\gamma_p \sim Beta(1, 1)$$
$$\theta_p \sim Dirichlet(\alpha_g^T \beta_g)$$
$$\nu_p \sim Beta(1 - \eta_g, \eta_g)$$

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

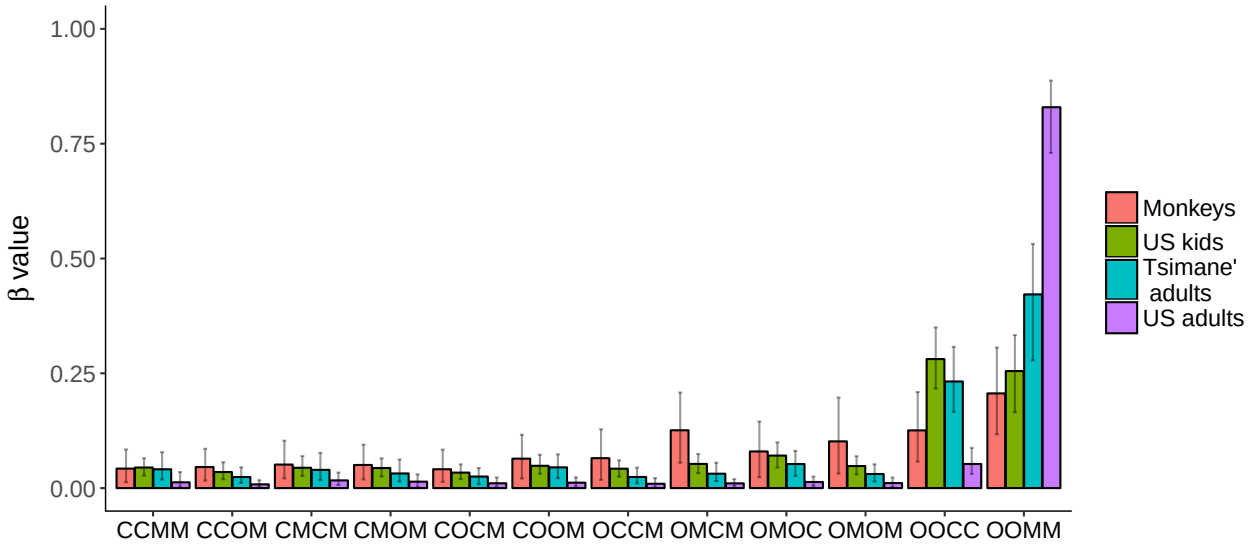
The function F, used to calculate R_p, adds noise and bias to the responses of strategies. Noise is added to each strategy's responses by determining every possible response's likelihood of having resulted from following that strategy. More specifically, responses that are more similar (have more overlap) 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 η is $\eta^D * (1 - \eta)^{4-D}$. 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 γ. So if γ 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

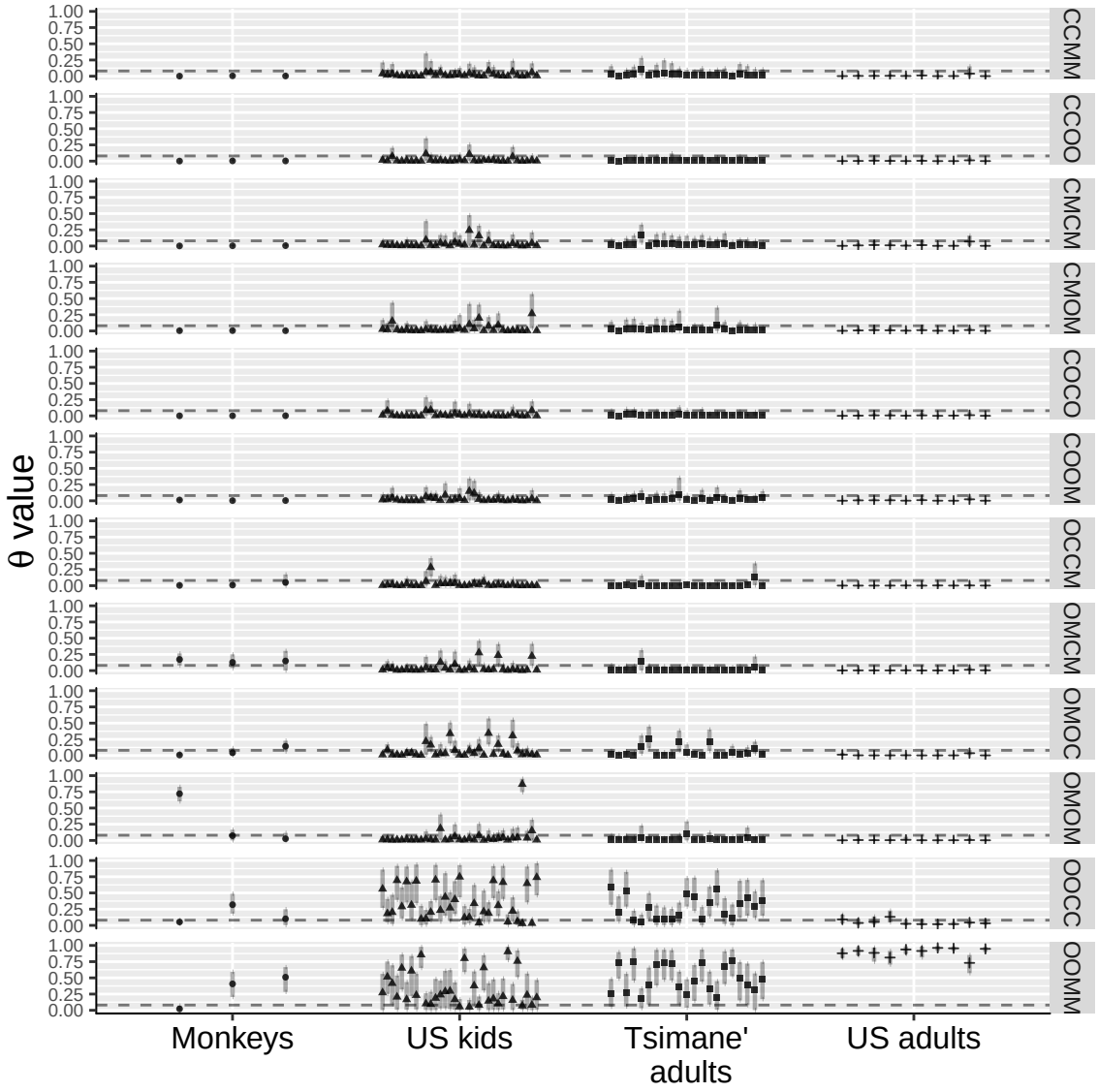
The Bayesian model was trained using PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016), with the default MCMC algorithm NUTS. It was run for 2,000 steps with 500 tuning steps, and a thin of 10. The low number of samples is due to NUTS being a gradient-based MCMC technique, and thus requires many fewer steps 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 deterermined by taking the smallest range containing 95% of samples.

Bayesian model results

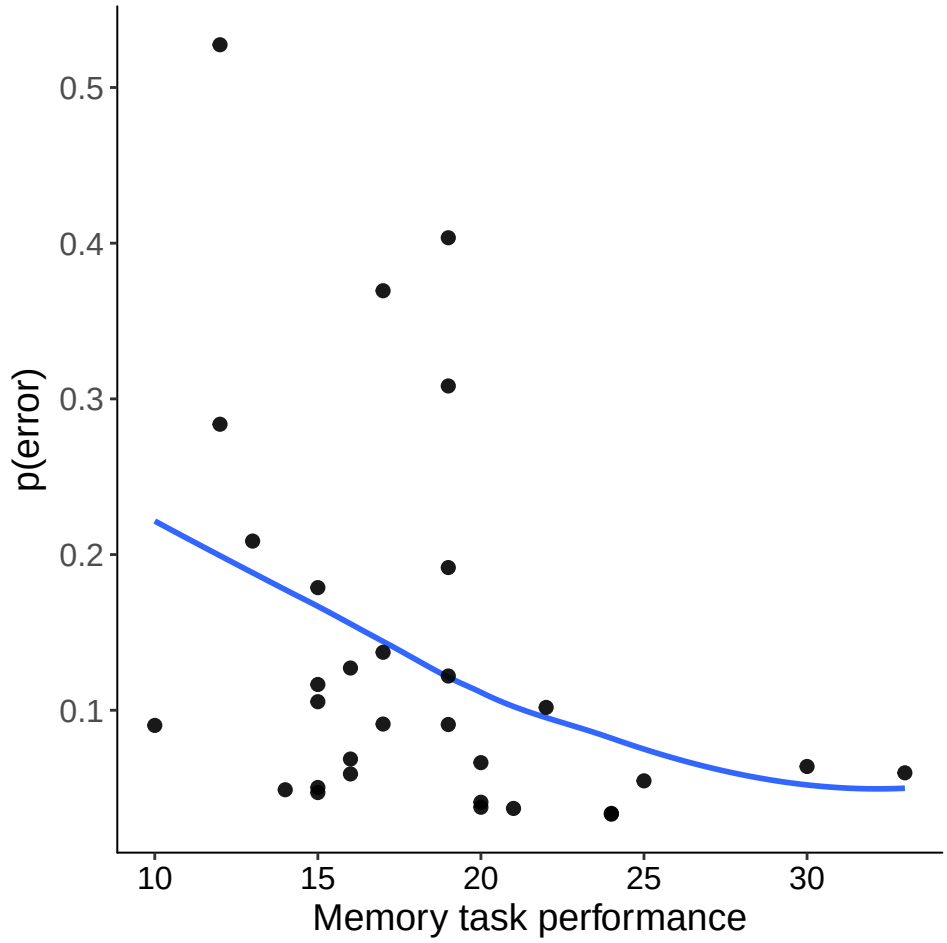
The full group-level means over strategies, represented by the parameter β_g, is shown in figure S2.



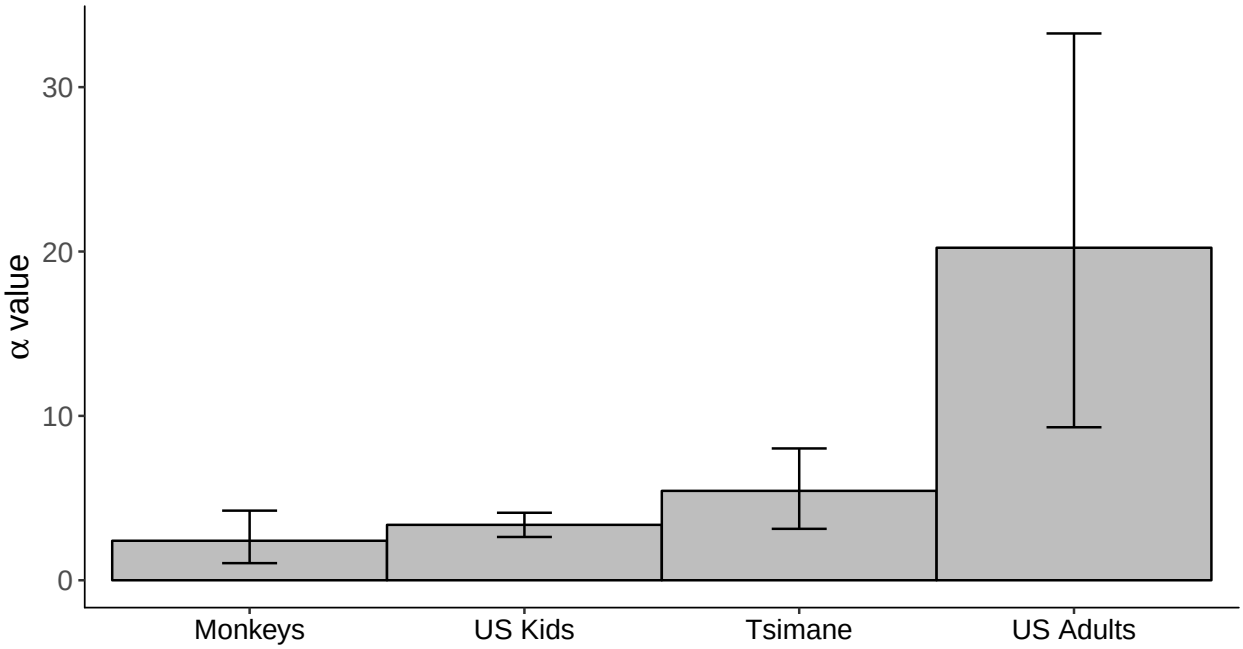
The inferred probabilities that individual participants were using each strategy, represented by the parameter θ_g, is shown in Figure S3.



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 ($\rho = -0.43,\ p = 0.02$). A plot of this effect is shown below.



There were differences between the group-level clustering parameter α. A higher α value corresponds to individuals in a group more tightly clustering around their group mean — so having more similar strategies. Adults had the highest α ($M = 9.2, CI = [4.5, 7.1]$), followed by monkeys ($M = 3.9, CI = [2.7, 4.8]$) and Tsimane ($M = 3.9, CI = [2.9, 4.5]$), and then kids ($M = 2.9, CI = [2.6, 3.1]$). It is intuitive that adults had the highest α — and thus were most tightly clustered — considering every adult was inferred to be using the strategy $OOMM$ with very high probability.


α value
30
20
10
0
Monkeys
US Kids
Tsimane
US Adults