Supplementary Materials for: Parsing Inhibitory and Mnemonic Contributions to Age-related Decline in Cognitive Flexibility

Emily R. Weichart, Cole S. Francis, and Per B. Sederberg

1 Supplementary Methods

1.1 Sample Size Justification

A pilot study of $N = 50$ participants was conducted on Prolific to estimate paired-samples effect sizes (d_z) of the main congruency and learning contrasts of interest. We acknowledge that simulation methods could have been used to conduct power analyses for the mixed effects models used in our primary analysis (Kumle et al., 2021). Because this work involves a novel experimental paradigm for which the underlying effects structure is not yet well-characterized, however, we selected an approach that would provide a reasonable sample size without requiring additional assumptions.

We subjected the pilot data to the same *a priori* inclusion/exclusion criteria that were applied to our main dataset. Twelve participants (24%) were excluded for failing to perform significantly above 50% accuracy, as determined by a binomial tests ($\alpha = 0.05$, $N_{trials} = 288$). An additional 2.3% of trials were excluded across participants for response times (RTs) that were faster than 200 ms or slower than 3500 ms. Among the remaining data from 38 participants, we observed a large effect ($d_z \approx 0.80$; **?**) of target array congruency on accuracy ($Accuracy_{congruent} - Accuracy_{180^\circ\ incongruent}$: $M = 0.091$, $SD = 0.100$, $t_{crit} = 2.201$, $d_z = 0.910$). A power analysis using *G*Power 3.1* (Faul et al., 2007) for a matched pairs difference between dependent means indicated that approximately 12 participants per age group would be needed to achieve 80% power (two-tailed, $\alpha = 0.05$). We additionally observed a very large effect ($d_z > 1.2$) of learning following a rule-switch ($Accuracy_{last\ 1/3\ of\ trials} - Accuracy_{first\ 1/3\ of\ trials}$: $M = 0.126$, $SD = 0.081$, $t_{crit} = 2.571$, $d_z = 1.556$). For this case, a power analysis indicated that 6 participants per age group would be needed to achieve 80% power (two-tailed, $\alpha = 0.05$). Taking the larger of these two estimates, we then made additional efforts to account for the age distribution of participants on Prolific as well as expected exclusions.

We selected age thresholds of 30 and 50 years to further inform a target sample size (young: ages 18-30; middle: ages 31-49; older: ages 50-73), such that a minimum of 12 participants were needed for each range. These thresholds were selected to align with prior studies that distinguish between cognitive profiles of younger (e.g. a university sample) and older adults (e.g. a community sample with age criteria of 50+),

although the group in between is typically excluded (Salthouse, 1994, 1996). Other age boundaries could have been selected to reflect developmental definitions of emerging adulthood, middle age, and older adulthood (Arnett, 2007; Maung, 2021). Within our pilot sample of $N = 50$ participants, $N = 18$ (36%) self-reported their age as being between 18 and 30, $N = 23$ (46%) were between 31 and 49, and $N = 9$ (18%) were 50 or older. For a target of 12 in the older adults group, we would thus need $12 \div 0.18 \approx 67$ participants to pass inclusion criteria. Given that 76% of our pilot sample passed performance-based inclusion checks, we expected to recruit a minimum of $67 \div 0.76 \approx 89$ participants. On this basis, we adopted a conservative recruitment target of $N = 100$ Prolific participants for the main study to ensure adequate representation across ages and to provide greater precision than the minimum estimated sample.

One-hundred (100) participants were pre-screened for English language fluency, normal or corrected-to-normal vision, and United States residency. All participants provided informed consent before beginning the experiment and were compensated after completion at a rate of 10 USD per hour (median completion time = 13.67 minutes). Sixteen participants (16%) were excluded from analyses for failing to exceed chance-level accuracy of 50% across trials, as determined by participant-wise binomial tests with $\alpha = 0.05$ and $N\,trials = 288$.

The final sample consisted of 84 participants (49 female, 34 male, 1 undisclosed sex) with a mean age of 38.3 years ($SD = 14.0$, range=18-73). Fifty-six (56; 66.7%) participants self-reported their race as White, 11 (13.1%) as Black, 10 (11.9%) as Asian, 6 (7.1%) as Mixed, and 1 (1.2%) as Other. In all statistical analyses reported here, age is used as a continuous measure. For descriptive purposes, 30 participants were aged 18-30 ($M = 24.7$), 35 were aged 31-49 ($M = 38.5$), and 19 were aged 50 and older ($M = 59.6$).

1.2 Task Training Module

Before beginning the experiment, participants completed a training module in an effort to ensure a thorough understanding of the learning environment. The module proceeded as follows:

1. Participants were first provided with a written description to introduce the task:

   > **Welcome!**
   >
   > In this game, you will see two arrows on the screen at a time. One arrow is the target, and the other is a distractor. Using the keyboard, you must press the arrow key that is pointing in the same direction as the target.
   >
   > Please position your right hand on the keyboard so that your forefinger is on the **LEFT** arrow key, your middle finger is on the **UP** arrow key, and your ring finger is on the **RIGHT** arrow key.

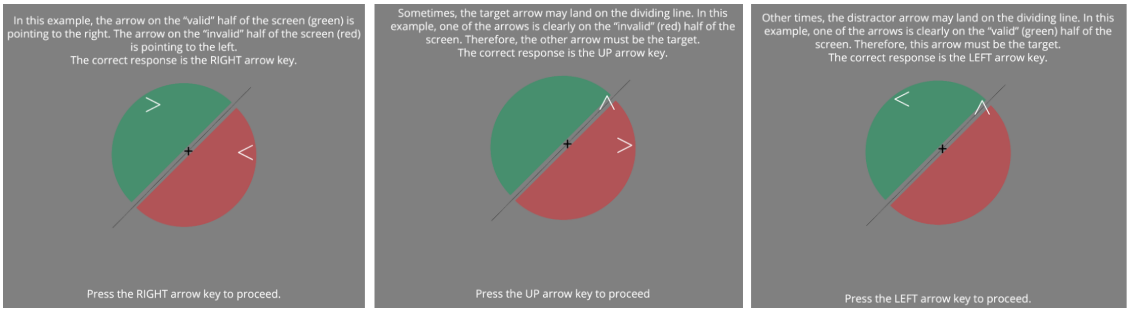
You will identify the target arrow by its location on the screen. The screen will be divided into two halves, where half of the screen contains "valid" locations and the other half contains "invalid" locations. The location of the target will change, but it will always appear on the half of the screen that is currently "valid."

2. Participants were then provided with three visual examples to illustrate the distinction between valid and invalid locations. The examples are shown in Figure S1. The following text was provided:

   A. In this example, the arrow in the "valid" half of the screen (green) is pointing to the right. The arrow in the "invalid" half of the screen (red) is pointing to the left. The correct response is the **RIGHT** arrow key (Figure 1A).

   B. Sometimes, the target arrow may land on the dividing line. In this example, one of the arrows is clearly on the "invalid" (red) half of the screen. Therefore, the other arrow must be the target. The correct response is the **UP** arrow key (Figure 1B).

   C. Other times, the distractor arrow may land on the dividing line. In this example, one of the arrows is directly on the "valid" half of the screen. Therefore, this arrow must be the target. The correct response is the **LEFT** arrow key (Figure 1C).



Supplementary Figure 1: Training example illustrating the distinction between valid (green) and invalid (red) halves of the screen. ***Panel A Text:*** In this example, the arrow on the "valid" half of the screen (green) is pointing to the right. The arrow on the "invalid" half of the screen (red) is pointing to the left. The correct response is the RIGHT arrow key. Press the RIGHT arrow key to proceed. ***Panel B Text:*** Sometimes, the target arrow may land on the dividing line. In this example, one of the arrows is clearly on the "invalid" (red) half of the screen. Therefore, the other arrow must be the target. The correct response is the UP arrow key. Press the UP arrow key to proceed. ***Panel C Text:*** Other times, the distractor arrow may land on the dividing line. In this example, one of the arrows is clearly on the "valid" (green) half of the screen. Therefore, this arrow must be the target. The correct response is the LEFT arrow key. Press the LEFT arrow key to proceed.

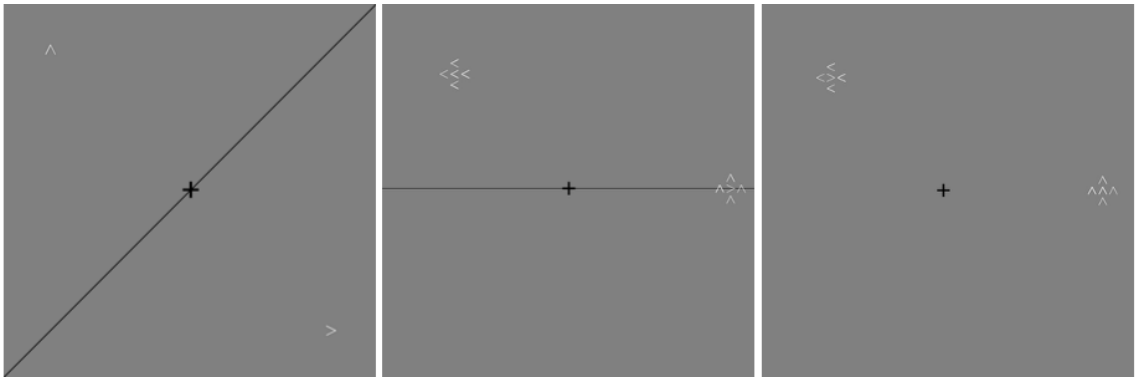
3. Participants then completed three practice sets, each containing 11 trials with correct/incorrect feedback. Instructions that preceded each practice set are provided below, and an example trial from each set is shown in Figure S2.

   A. *Let's practice.* Use the arrow keys to indicate which direction you think the target arrow is pointing. After you respond, we will tell you if you were correct or incorrect. You will have to

use trial-and-error to figure out which side of the line is "valid," and which is "invalid." (Figure 2A)

B. In the next practice set, you are going to see two arrays of arrows. Use the arrow keys to indicate which direction you think the arrow in the **CENTER** of the target array is pointing. Just like before, you will have to use trial-and-error to figure out which side of the line is "valid," and which is "invalid." (Figure 2B)

C. This time, you will not see the line that divides the screen into "valid" and "invalid" areas. Just as before, use the arrow keys to indicate which direction you think the **CENTER** arrow of the target array is pointing. (Figure 2C)



Supplementary Figure 2: Example trials from the three practice sets in the training module.

4. Finally, participants read a summary of the instructions before proceeding to the main experiment:

   Now we will begin the real game!
   Use the **LEFT**, **UP**, and **RIGHT** arrow keys to indicate which direction the center arrow of the target array is pointing. Just like in the last practice set, you will not see the line that divides the "valid" from the "invalid" parts of the screen. You must use trial-and-error to figure out which parts of the screen are "valid" in order to successfully identify the target array. The parts of the screen that are "valid" and "invalid" will change at unannounced times during the game. Even if you make mistakes, do your best to respond as quickly and accurately as you can.

2 Supplementary Results

2.1 Error Types

To further clarify the results of the first accuracy analysis, we examined the types of errors participants tended to make when responding to incongruent arrays. On each trial, "left," "right," and "up" response options

corresponded to distinct elements of every trial-level stimulus, allowing us to infer whether the participant's response aligned with the target or distractors, and whether the corresponding arrow came from the valid or invalid array. Relative to the primary accuracy analysis, focusing on incongruent arrays resulted in an exclusion of 49.8% of trials across all participants. We used Bayesian multinomial logistic regression with a softmax link to model categorical responses with the following levels: 1) valid array, target (correct; reference category); 2) valid array, distractors (error); 3) invalid array, target (error); and 4) invalid array, distractors (error). Predictors were age (z-scored), the log-transformed trial number following a rule switch (last trial as reference level), and their interaction. By-subject intercepts were included as random effects. The resulting coefficients and highest density intervals (HDIs) are provided in Table 1. Asterisks indicate HDIs that exclude 0, which we consider to indicate predictor credibility.

Supplementary Table 1: Coefficients from error types analysis, using multinomial logistic regression.

Predictor	**Mean**	**SD**	**94% HDI**
Intercept[invalid-target]	-2.768	0.094	[-2.946, -2.594]*
Intercept[valid-flankers]	-3.281	0.122	[-3.505, -3.045]*
Intercept[invalid-flankers]	-5.349	0.200	[-5.711, -4.959]*
Trial num. [invalid-target]	-0.606	0.025	[-0.653, -0.562]*
Trial num. [valid-flankers]	0.141	0.063	[0.025, 0.261]*
Trial num. [invalid-flankers]	-0.326	0.097	[-0.514, -0.155]*
Age [invalid-target]	0.328	0.091	[0.159, 0.506]*
Age [valid-flankers]	-0.479	0.133	[-0.729, -0.233]*
Age [invalid-flankers]	0.537	0.197	[0.155, 0.896]*
Age $\times$ Trial num. [invalid-target]	0.112	0.024	[0.066, 0.158]*
Age $\times$ Trial num. [valid-flankers]	-0.068	0.071	[-0.201, 0.065]
Age $\times$ Trial num. [invalid-flankers]	0.399	0.107	[0.198, 0.593]*

2.2 Block-level Effects

To examine the possibility of practice effects of effects of fatigue throughout the task, we re-ran our primary analysis to include interactions with block number. Predictors were age (z-scored), the log-transformed trial number following a rule switch (last trial as reference level), trial type of the valid array (congruent as reference level), and block number (first block as reference level) and their interactions. By-subject intercepts were included as random effects. Results are provided in Table 2. We observe that end-of-block accuracy on congruent trials decreases across blocks (p=0.014). Additionally, while accuracy on 90$\circ$ incongruent trials is lower than congruent in the first block (p=0.002), this performance difference is reduced in later blocks (p=0.009). Notably, no interactions involving age and block number are significant and all main effects from our primary analysis appear to be robust to limiting the model's baseline to the first block.

Supplementary Table 2: Extended logistic regression including interactions with block number.

Predictor	Estimate	SE	Z	p
(Intercept)	2.743	0.091	30.010	$< .001^{***}$
Age	-0.226	0.092	-2.464	$.014^{*}$
Trial type (90° inc.)	-0.233	0.044	-3.047	$.002^{**}$
Trial type (180° inc.)	-0.562	0.088	-6.348	$< .001^{***}$
Trial number	0.566	0.030	18.742	$< .001^{***}$
Block number	-0.040	0.016	-2.452	$.014^{*}$
Age × Trial type (90° inc.)	0.078	0.078	1.003	.316
Age × Trial type (180° inc.)	0.276	0.092	3.002	$.003^{**}$
Age × Trial number	-0.064	0.031	-2.092	$.036^{*}$
Age × Block number	-0.001	0.016	-0.046	.964
Trial type (90° inc.) × Block number	0.048	0.018	2.620	$.009^{**}$
Trial type (180° inc.) × Block number	0.015	0.021	0.712	.477
Trial number × Block number	-0.002	0.007	-0.261	.794
Age × Trial type (90° inc.) × Block number	-0.015	0.018	-0.793	.428
Age × Trial type (180° inc.) × Block number	-0.026	0.021	-1.242	.214
Age × Trial number × Block number	0.003	0.007	0.477	.633

3 Supplementary Computational Modeling

3.1 Fixed Values and Prior Specification

The model consists of 5 parameters: α_{flank} (attention to distractors), δ (memory precision), λ (task-level inhibition), δ_p (perceptual precision), and θ (determinism of responding). In order to investigate specific questions regarding the relationships between age and mechanisms relevant to learning and inhibition, δ_p and θ were fixed across participants. Their values ($\delta_p = 6$ and $\theta = 3$) were selected via a grid search approach that aimed to maximize total log-likelihood across participants. This entailed: 1) systematically sampling the prior ranges of δ_p and θ; 2) fixing the values and fitting the model to data; 3) calculating the total log-likelihood across participants. Priors for the remaining free parameters were specified by sampling from a standard Normal distribution ($x \sim Normal(0,1)$), and applying an inverse logit transformation:

$$x \rightarrow R \cdot \frac{e^x}{1+e^x} \tag{1}$$

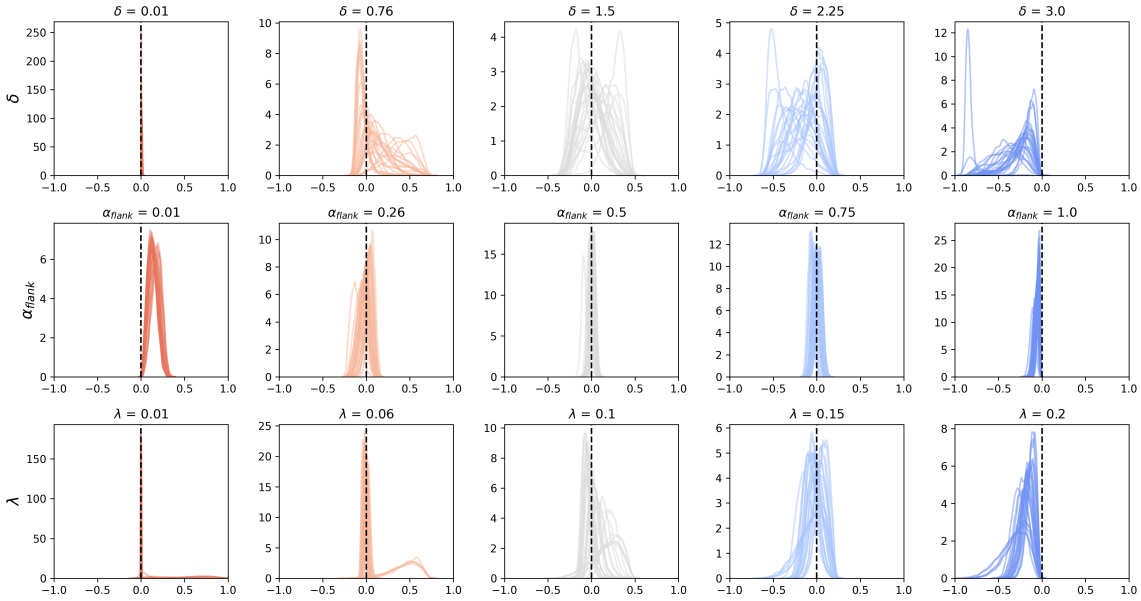
where R sets the upper bound of the parameter's effective range. This approach resulted in weakly informative priors over the range $(0,R)$ that were centered at $\frac{R}{2}$ and avoided boundary truncation.

Values of R were selected based on theoretical constraints and model behavior. For α_{flank}, $R = 1$ was selected to be consistent with prior exemplar model specifications where $\sum \alpha = 1$ (Nosofsky, 1986). This constraint was satisfied in our model, given that $\alpha = [\alpha_{targ}, \alpha_{flank}]$ and $\alpha_{targ} = 1 - \alpha_{flank}$. For δ, we set $R = 3$ given that higher values yielded negligible additional impact on model predictions with $\theta = 3$ and we wanted to ensure that the parameter remained interpretable on a behavioral scale. For λ, we set $R = 0.2$ to capture a region where inhibition modulated learning in a linearly-interpretable way. Whereas values below

0.2 produced perseveration-like behaviors of interest, values above 0.2 yielded immediate inhibition of all previous information unless δ was simultaneously increased.

3.2 Parameter Recovery

A parameter recovery study confirmed that all three free parameters were identifiable over their respective prior ranges. We selected 5 evenly-spaced values across each parameter's prior (including the endpoints), and identified all possible combinations of values across the 3 parameters. This resulted in a full grid of 125 parameter sets. For each parameter set, we generated behavior using the model and the trial sequence (288 trials) from a randomly-selected participant. The model generated a valid probability distribution over "left," "right," and "up" responses for each trial, from which a single response was drawn using multinomial sampling. The model was then fit to each simulated dataset, and posterior means were extracted for each parameter. Recovery was evaluated by comparing the recovered posterior means to the true generating values across the entire grid (Figure 3). The Pearson correlations between true and recovered values indicated good recovery, insofar as $\rho > 0.8$ for all parameters (Danwitz et al., 2022): ($\alpha_{flank} : \rho = 0.99$, $\delta : \rho = 0.85$, $\lambda : \rho = 0.87$).



Supplementary Figure 3: Aggregate differences between generating parameters (centered at 0) and recovered posterior distributions

3.3 Prediction Validation

Before we could justifiably use parameter estimates to make inferences about age-related differences in cognitive mechanisms, it was necessary to verify that the computational model could capture the behavioral effects observed in our primary analysis. After fitting the model to each participant's data, we extracted the

subject-level posterior mean for each parameter. Using these values, we generated data by presenting the model with the exact sequence of stimuli shown to each respective participant. This produced trial-by-trial predicted response probabilities for each of the three choice options. To generate simulated responses, we then performed a single multinomial draw on each trial based on these probabilities. This process was repeated 10 times per participant to generate 840 nested replicate datasets. We used this approach to 1) generate binary outcomes so that we could apply the same logistic mixed effects regression model that we used in our primary analysis; and 2) 10 replicate datasets per participant provided stable coefficient estimates and reduced the influence of randomness in any single set of multinomial draws. For methodological consistency with the generative structure of the computational model, we subjected the replicate datasets to logistic mixed effects regression within a Bayesian framework.

Supplementary Table 3: Logistic mixed effects regression coefficients for predicting model-generated accuracy.

Predictor	**Observed Data**			**Model-Simulated Data**		
	Mean	**SD**	**94% HDI**	**Mean**	**SD**	**94% HDI**
Intercept	2.596	0.072	[2.458, 2.731]	2.585	0.051	[2.488, 2.680]
Age	-0.232	0.071	[-0.364, -0.097]	-0.202	0.049	[-0.291, -0.107]
Trial type (90° inc.)	-0.064	0.043	[-0.142, 0.016]	-0.125	0.013	[-0.148, -0.100]
Trial type (180° inc.)	-0.507	0.048	[-0.598, -0.419]	-0.521	0.015	[-0.548, -0.494]
Trial number	0.559	0.016	[0.529, 0.590]	0.630	0.005	[0.621, 0.640]
Age × Trial type (90°)	0.029	0.042	[-0.051, 0.107]	-0.007	0.013	[-0.030, 0.017]
Age × Trial type (180°)	0.182	0.049	[0.088, 0.270]	0.150	0.015	[0.121, 0.178]
Age × Trial number	-0.052	0.017	[-0.084, -0.022]	-0.051	0.005	[-0.061, -0.042]

3.4 Mediation Analysis

We conducted mediation analyses to examine whether the observed age-related decrease in flexible learning could be explained by changes in memory precision (δ) or set-level inhibition (λ). As described by Baron & Kenny (1986), we observe a total effect of an independent variable X (here, age) on a dependent variable Y (here, some measure of performance), as represented the regression equation:

$$Y = \beta_1 + cX + \varepsilon_1 \tag{2}$$

where β_1 is the intercept, ε_1 is the error, and coefficient c is significantly different from zero. We may additionally observe an effect of age on one of our computational model parameters M, such that coefficient a is significant in the regression equation:

$$M = \beta_2 + aX + \varepsilon_2. \tag{3}$$

To examine the effect of age X on performance outcome Y while controlling for parameter M, we specify the equation:

$$Y = \beta_3 + c'X + bM + \varepsilon_3. \tag{4}$$

The *indirect effect* of X on Y via M is the product of coefficients a (Equation 2) and b (Equation 3). The *direct effect* of X on Y controlling for M is thus c' (Equation 4). If the indirect path (ab) is significant while reducing the influence of the direct path (c') to zero, we would conclude that the relationship between X and Y is *fully mediated* by M. If we find that both the indirect and direct paths are significant, we would conclude that the relationship between X and Y is *partially mediated* by M. If the indirect path is insignificant while the direct path remains significant, we would conclude that the relationship between X and Y is *not* mediated by M.

Among the combinations of M and Y where $Y \sim X$ and $Y \sim M$, the third step was to apply the methods of Yuan & MacKinnon (2009) for single-level Bayesian mediation. This entailed two linear regression models: $Model\ i$ predicted parameter M from age, and $Model\ ii$ predicted performance outcome Y from age and parameter M. Using Figure **??** for reference, a was defined as the coefficient posterior when predicting M from age in $Model\ i$. b was the coefficient on M when predicting Y in $Model\ ii$. Finally, c' was the coefficient on age when predicting Y in $Model\ ii$. The indirect (mean of $a \times b$) and the direct (mean of c') paths were considered to have credible impacts on Y if their respective 94% HDIs excluded zero.

References

Arnett, J. (2007). Emerging adulthood: What is it, and what is it good for. *Child Development Perspectives*, *1*(2), 68–73. doi: {https://doi.org/10.1111/j.1750-8606.2007.00016.x}

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. doi: {https://doi.org/10.1037/0022-3514.51.6.1173}

Danwitz, L., Mathar, D., Smith, E., Tuzsus, D., & Peters, J. (2022). Parameter and model recovery of reinforcement learning models for restless bandit problems. *Computational Brain & Behavior*, *5*(1), 547–563. doi: https://doi.org/10.1007/s42113-022-00139-0

Faul, F., Erdfelder, E., & Lang, A.-G. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi: https://doi.org/10.3758/BF03193146

Kumle, L., Vo, M., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. (https://doi.org/10.3758/s13428-021-01546-0)

Maung, H. (2021). What's my age again? Age categories as interactive kinds. *History and Philosophy of the Life Sciences*, *43*(36). doi: {https://doi.org/10.1007/s40656-021-00388-5}

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi: {https://doi.org/10.1037//0096-3445.115.1.39}

Salthouse, T. (1994). Aging associations: Influence of speed on adult age differences in associative learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1486–1503. doi: {https://psycnet.apa.org/doi/10.1037/0278-7393.20.6.1486}

Salthouse, T. (1996). General and specific speed mediation of adult age differences in memory. *The Journals of Gerontology: Series B*, *51B*(1), 30–42. doi: {https://doi.org/10.1093/geronb/51B.1.P30}

Yuan, Y., & MacKinnon, D. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301–322. doi: https://psycnet.apa.org/doi/10.1037/a0016972