Round #2

by Aurélie Coulon, 14 Mar 2023 10:29
Manuscript: https://doi.org/10.32942/X2R59K version 2
revision needed for the preprint "Using repeatability of performance within and across contexts to validate measures of behavioral flexibility"

Dear Dr McCune and collaborators,

I have received two reviews of the revised version of your preprint called “Repeatability of performance within and across contexts measuring behavioral flexibility” (now renamed “Using repeatability of performance within and across contexts to validate measures of behavioral flexibility”). They were written by the same reviewers who evaluated the first version of this preprint. Both of them are satisfied with the way you dealt with their previous comments. One of them, Maxime Dahirel, still has one important comment that needs to be addressed (and a few more minor points). Once these comments are addressed, your preprint should be ready for recommendation.

Best,

Aurélie Coulon.
Recommender, PCI Ecology.

Reviews

Reviewed by Maxime Dahirel, 16 Feb 2023 12:06
I have now read the revised version of “Repeatability of performance within and across contexts measuring behavioral flexibility” by McCune et al.

I have mostly good things to say about this revision. The authors took into account most of my comments, and when they did not, they provided enough justification both in the reply and in the revised manuscript that even if I may have wished for different changes, I am OK enough with their choices as is, given the scope of the manuscript. I have only one major comment remaining: the authors treat their results for within-context repeatability as clear evidence that this repeatability is different from 0. The situation is actually a bit more nuanced and muddy; see my comment 1 below. Beyond that, my remaining comments are mostly minor and trivial to reply to, I think. As long as the authors keep in mind the limitations of their dataset in the Discussion (which they already do fairly well), the outcome will be in my opinion a clear, short and interesting manuscript.

COMMENT 1a: I have a bit of issue with discussing the temporal repeatability estimate as unambiguously “different from 0”, since in actuality it might not be.

- The lower bound of the 95% CI, 4.10^-16 is functionally == 0, especially in a Bayesian modelling context where generally, estimates of variance components can never be exactly == to zero, contrary to frequentist ones. (not always, it may depend on priors and exact implementation details of course, but in most default cases). Therefore, whether we can say that our estimate is meaningfully different from 0 will depend on the shape of the whole posterior, and on how high is our threshold to say something is not == 0. In practice, I am OK with a narrative saying that individual identity explains on average 13% of variance, but that uncertainty remains large and there is still the possibility it explains 0, and that puts this value in context with e.g. other estimates of behavioural repeatability in the literature.

Related to that: in Figure 3, as far as I understand it, the method compares observed average repeatability (in red) to simulated average repeatabilities (the histogram). Fair enough, but this is missing that both these repeatabilities have substantial uncertainties that are ignored by this method, which means I am not sure at all this comparison is meaningful at all. Indeed, I can easily imagine a dataset that matches all the characteristics of yours, gives this figure when analysed, and yet the underlying observed repeatability would not be different from 0. The "correct" comparison would involve comparing the entire posteriors I think, not the posterior means.

Response 1a: First we want to say thank you for again taking the time to review our revised manuscript and for thinking critically about how it can be improved.

We agree that the repeatability estimate is interpreted as the amount of variance in the data that can be attributed to individual identity. But, the question then becomes, is that amount of variance statistically significant? In Nakagawa & Schielzeth (2010), the recommended method for this is not to interpret the confidence/credible intervals, but instead to use likelihood ratio tests or permutation tests to evaluate the significance of the individual ID random effect. Given the constraints of the serial reversal learning protocol, where the end of the experiment is determined when an individual passes two reversals in 50 or fewer trials, we used the latter permutation approach to evaluate whether 0.13 is greater than repeatability estimates if birds were performing randomly.

You are right that Figure 3 in the manuscript showed the shape of the average of posteriors from each simulation of the *permuted* data. In response to your comment, we investigated the shape of all of the posterior estimates from the actual and permuted data to evaluate the uncertainty in each set of estimates. We randomly sampled, with replacement, from the values of the posterior of the model using the actual data until we had the same sample size as for the permuted data (Fig. R1, below). We then looked at the probability that one posterior estimate of repeatability from our actual data would be greater than a sample posterior from the permuted data. 79% of the time, a randomly selected estimate from the actual data was greater than the estimate from the permuted data that it was paired with. This indicates that values from the full posterior distribution primarily do not overlap. Moreover, the mean of the permuted data is 0.008, whereas the mean of our actual data is 0.13, as reported in the manuscript. Thus, we can be confident that the mean repeatability in our actual data is different from random.

To address your comment, we added to the text a new Fig. 3 (the same as Fig. R1, seen below) that shows that even though the lower bound of the actual data is close to 0, the majority of the posterior values are greater than 0 and greater than the permuted posterior estimates.

And we wrote additional details in the Results section: “We found that, although the lower bound of the credible interval is approximately zero, the mean repeatability value was significantly greater than expected if birds were performing randomly [p=0.003; @nakagawa2010repeatability]. Furthermore, the distribution of the posterior estimates for the actual data were much less skewed towards zero compared to the permuted data of birds performing randomly (Fig. 3; see analysis details in the R code for Analysis Plan > P3a), though with the uncertainty we cannot completely exclude that individual identity might not influence performance.”

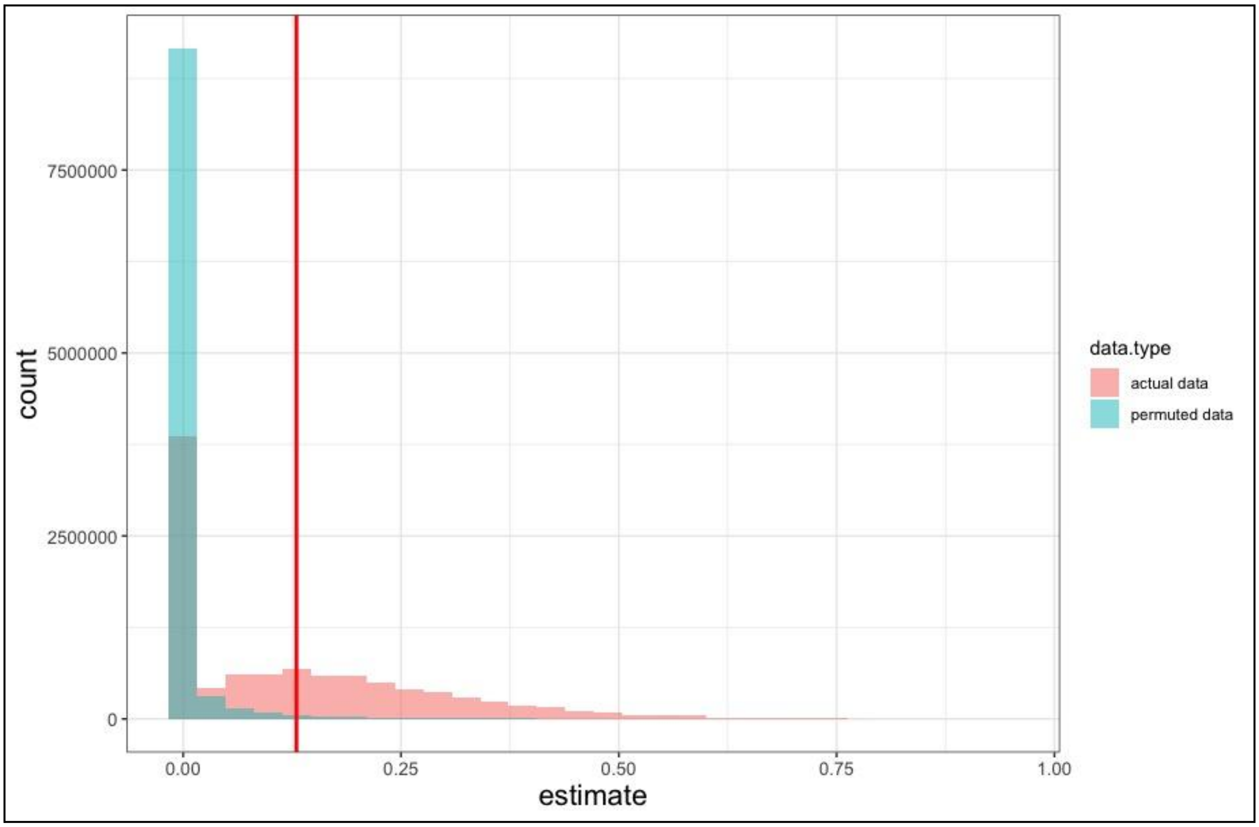

Figure R1: Frequency histograms of posterior repeatability estimates from the model testing the latency to switch a color preference in a serial reversal learning experiment. To determine the significance of our repeatability value while accounting for the non-independence of the serial reversal learning experimental design, we compared our repeatability value to repeatability values calculated from simulated data where birds performed randomly within each reversal. Estimates from actual data (red) are compared to the distribution of estimates from randomized

permutations of the data (green). The vertical red line at 0.13 is the observed mean repeatability estimate reported in this manuscript and it was significantly greater than random.

Comment 1b: Note that this does not mean that temporal repeatability is inexistant. Indeed I think there is some interesting discussion to have about the fact there may be stronger evidence in your data for between-context vs temporal repeatability (but again, see huge uncertainties and small sample sizes). How could a trait be repeatable between contexts if it's not within context? The timeline of the tests here might be very important (intervals between tests, from the same experiment and different experiments). Repeatabilities are expected to differ between time scales (see e.g. Harrison et al., 2019).

Response 1b: While repeatability of behavior in some cases is dependent on the time between samples (Harrison et al. 2019; Bell et al. 2009), other reviews on the repeatability of cognitive performance found no effect of time between samples (Cauchoix et al. 2018). In our study, all individuals received the serial reversal learning experiment before the multiaccess box experiments. After they passed serial reversal learning, the subsequent tests were given in a randomized order across subjects. Consequently, subjects received a multiaccess box task between 0 and 142 days after completing reversal learning because there was a lot of variability in how long it took subjects to complete each reversal. So, in this study it was impossible to control for the time between samples within or across contexts. But this would be interesting to evaluate in a future study where time between samples is specifically manipulated.

Because our repeatability estimate was larger for data on performance across contexts compared to within contexts, we agree there could be interesting aspects to further investigate there. We added a new paragraph to the Discussion on this point:

"The repeatability estimate for cross-contextual switching performance was higher than the estimate for switching performance within a context, indicating that a larger portion of the variance in cross-contextual performance is attributable to individual differences (lower residual variance and/or greater among-individual variance). Performance on a task likely depends on multiple cognitive processes, some of which might be more repeatable than others. For example, @lukas2022flexmanip found that performance on the serial reversal learning task was related to two distinct components - the rate of updating an attraction to a colored tube (phi) and the likelihood of deviating from the learned attractions (lambda), where phi appeared to show more individual consistency than lambda. Repeatability might be higher for cross-contextual switching depending on which cognitive processes dominate in a given task and across contexts. Variation in the design of our tasks may lead to higher residual variance in individual performance across reversals because food is hidden in the serial reversal learning task but clearly visible behind transparent plastic barriers in both MAB tasks. After a reversal, to determine which of the two colored tubes to search in for food, grackles cannot rely on short term memory of the previous location of food, they must have some motivation to search in a new color of tube (lambda). Consequently, it is possible that higher within-individual variation in performance across serial reversals in the latency to switch was related to the other factors affecting an individual's decision-making on each trial, like conflicting memories of reward history for each color tube or a tendency to make a choice based on a side bias. In contrast, in the MAB tasks, even if the previously rewarded option is non-functional, the grackles can clearly see that the food is still there, which may facilitate motivation to change their behavior regardless of past memories of reward contingencies or bias towards certain stimuli."

Comment 2: Lines 56-74: This is a very good overview of how important repeatability/consistent inter-individual variance is as a criterion to decide if a trait is a valid proxy of cognition, but doesn't address a key limitation in my opinion: it's not actually "latent persistent cognition" vs "ephemeral factors" shaping cognitive proxies, it's "latent cognition", "latent persistent other traits", and "ephemeral factors". Persistence, motivation, physiology et al. can all be repeatable to some extent... and assessing the repeatability of the behavioural task in general is the first step only. This relates to some comments on the first version, and I would appreciate to see some discussion of this limitation at least in the discussion.

Response 2: You are right that even with temporal and contextual repeatability, performance could still reflect other traits in addition to the one of interest. But it does greatly increase the probability of pinpointing specific cognitive traits affecting performance. So a goal of our manuscript was to point out that this is one important step to make because many researchers are not yet incorporating repeatability for convergent and divergent method validity into animal cognition studies. Other studies in highly controlled environments are more able to control for ephemeral factors and other latent consistent traits, but the temporal and contextual repeatability methods we demonstrate here are feasible for researchers working on wild animals in their natural environment as well as in captivity.

To respond to your comment, we toned down our assertion on this point in the Discussion:

First paragraph: "...This indicates that 1) the different methods we used to measure behavioral flexibility all likely measure the same **latent inherent aspects affecting performance**…"

Second paragraph: "...Although our sample size was small, which likely led to moderately large credible intervals, we still found significant temporal and contextual repeatability of switching performance. This evidence for convergent validity indicates these similar tasks are likely assessing **aspects of the same latent trait or traits** [@volter2018comparative; @morand2022cognitive]. **However, performance can potentially be affected by many traits, so future studies manipulating other factors that might influence performance are needed to continue to pinpoint the latent traits governing aspects of performance on cognitive tasks. Thus,** it is important to also test for discriminant validity by comparing performance on flexibility tasks with tasks intended to measure different cognitive abilities. For example, it is possible that performance on serial reversal learning and solution switching on the

MAB tasks is reflective of a trait other than behavioral flexibility, like inhibition [@maclean2014evolution]...”

Comment 3: Line 193-195: This definition is wrong. The authors are implying here, or even outright writing, that a trait is repeatable if there is more among- than within-individual variance (so if r > 0.5). No, a trait is repeatable if there is detectable among-individual variance, it doesn’t have to be higher than within-individual variance. Given the typical behavioural repeatability is <0.5 (see e.g. Holtmann et al., 2017), the authors’ phrasing would imply we need to discard a lot of the literature, including actually their own results in the present study.

Response 3: Thank you for catching this, it is an error in wording on our part. As you suggest, we updated this sentence to read:

“In other words, performance is likely to represent an inherent trait when there is significant among-individual variation in performance across repeated samples.”

Comment 4: Line 202-204: this sentence implies the authors used adjusted repeatabilities (Nakagawa & Schielzeth, 2010), ignoring the variance linked to the fixed effect in your denominator. That’s perfectly OK if that’s what’s intended, but that needs to be stated explicitly.

In addition, I’m using this comment to point that the likelihood/family of the model should be specified explicitly here. We shouldn’t need to wait until the detailed prereg/deviations from prereg to know that the analysis was done using a Gaussian LMM with potentially transformed data; this hurts readibility.

Response 4: We apologize that this was unclear. We did use adjusted repeatability to account for variance attributed to the fixed effect (which was reversal number in this case). It seems there may be a typo in your comment because Nakagawa & Schielzeth (2010) define adjusted repeatability as repeatability after controlling for confounding effects of fixed effects (p 937 & 949). We clarified this, and the family of the model we used, in the text:

“We thus used the adjusted repeatability [@nakagawa2010repeatability] as the variance components for the random effect and residual variance, after accounting for the variance attributed to reversal number, to determine the proportion of variance attributable to differences among individuals. Although our dependent variable (number of trials to reverse) is a count variable, the distribution of values was not appropriate for a poisson regression. When checking the fit of our data to a poisson model, the data were overdispersed and heteroscedastic. However, when log-transformed, the data approximate a normal distribution and are not heteroscedastic, indicating the Gaussian model fits our log-transformed data well.”

Comment 5: Lines 46-49: “captive animals” instead of “captive individuals”? I suggest that choice because the sentence include mentions of humans earlier, and a cursory reading/tired reader may think for a second we’re talking about captive humans here.

Response 5: Yes, good point! That is a funny and unfortunate misreading that could occur. We changed this word to animals as you suggested.

Comment 6: Line 286: 9 individuals, but how many data points total? please precise both, both are important for understanding the estimates.

Response 6: This data set included 9 individuals with one value for each reversal they experienced. Grackles experienced between 6 and 11 reversals. Therefore there are 68 total data points. We clarified this in the text (changes in bold face font):

“Our sample size was 9 **individual grackles and 68 total data points (one value for each of the 6-11 reversals that each grackle experienced)** for our first hypothesis testing temporal repeatability of reversal learning performance.”

Comment 7: Figure 4: “lines indicate the variation”: is it the range of values? IQR? SD? SE? This should be explicit.

Response 7: Apologies for the lack of clarity here. The lines are the interquartile range. We edited this figure caption as follows:

“... the lines indicate **the interquartile range of** variation in performance…”

REFERENCES

Harrison, P. M., Keeler, R. A., Robichaud, D., Mossop, B., Power, M., & Cooke, S. J. (2019). Individual differences exceed species differences in the movements of a river fish community. Behavioral Ecology, arz076. https://doi.org/10.1093/beheco/arz076

Holtmann, B., Lagisz, M., & Nakagawa, S. (2017). Metabolic rates, and not hormone levels, are a likely mediator of between-individual differences in behaviour: A meta-analysis. Functional Ecology, 31(3), 685–696. https://doi.org/10.1111/1365-2435.12779

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. Biological Reviews, 85(4), Article 4. https://doi.org/10.1111/j.1469-185X.2010.00141.x

Reviewed by Aparajitha Ramesh, 09 Mar 2023 21:07

I thank the authors for addressing all my comments and concerns. I find that the flow and focus of the paper has improved compared to the previous version. I believe this version of the manuscript to be a valuble addition to the current literature on animal cognition and animal personality.

Response: We again give our thanks for your efforts reviewing our manuscript. We are glad that we satisfied your comments and concerns with our revision and we believe the manuscript is much improved.