Peer Community In Ecology

Do reversal learning methods measure behavioral flexibility?

Aurélie Coulon based on peer reviews by *Maxime Dahirel* and *Aparajitha Ramesh*

McCune KB, Blaisdell AP, Johnson-Ulrich Z, Lukas D, MacPherson M, Seitz BM, Sevchik A, Logan CJ (2023) Using repeatability of performance within and across contexts to validate measures of behavioral flexibility. EcoEvoRxiv, ver. 5, peer-reviewed and recommended by Peer Community in Ecology. https://doi.org/10.32942/X2R59K

Submitted: 15 August 2022, Recommended: 26 May 2023

Cite this recommendation as:
Coulon, A. (2023) Do reversal learning methods measure behavioral flexibility?. *Peer Community in Ecology*, 100467. 10.24072/pci.ecology.100467

Published: 26 May 2023

Assessing the reliability of the methods we use in actually measuring the intended trait should be one of our first priorities when designing a study – especially when the trait in question is not directly observable and is measured through a proxy.

This is the case for cognitive traits, which are often quantified through measures of behavioral performance. Behavioral flexibility is of particular interest in the context of great environmental changes that a lot of populations have to experiment. This type of behavioral performance is often measured through reversal learning experiments (Bond 2007). In these experiments, individuals first learn a preference, for example for an object of a certain type of form or color, associated with a reward such as food. The characteristics of the rewarded object then change, and the individuals hence have to learn these new characteristics (to get the reward). The time needed by the individual to make this change in preference has been considered a measure of behavioral flexibility.

Although reversal learning experiments have been widely used, their construct validity to assess behavioral flexibility has not been thoroughly tested. This was the aim of McCune and collaborators' (2023) study, through the test of the repeatability of individual performance within and across contexts of reversal learning, in the great-tailed grackle.

This manuscript presents a post-study of the preregistered study* (Logan et al. 2019) that was peer-reviewed and received an In Principle Recommendation for PCI Ecology (Coulon 2019; the initial preregistration was split into 3 post-studies).

Using 34 great-tailed grackles wild-caught in Tempe, Arizona (USA), the authors tested in aviaries 2 hypotheses:

- First, that the behavioral flexibility measured by reversal learning is repeatable within individuals across sessions of the same experiment;
- Second, that there is repeatability of the measured behavioral flexibility (within individuals) across different types of reversal learning experiments (context).

The first hypothesis was tested by measuring the repeatability of the time needed by individuals to switch color preference in a color reversal learning task (colored tubes), over serial sessions of this task. The second one was tested by measuring the time needed by individuals to switch solutions, within 3 different contexts: (1) colored tubes, (2) plastic and (3) wooden multi-access boxes involving several ways to access food.

Despite limited sample sizes, the results of these experiments suggest that there is both temporal and contextual repeatability of behavioral flexibility performance of great-tailed grackles, as measured by reversal learning experiments.

Those results are a first indication of the construct validity of reversal learning experiments to assess behavioral flexibility. As highlighted by McCune and collaborators, it is now necessary to assess the discriminant validity of these experiments, i.e. checking that a different performance is obtained with tasks (experiments) that are supposed to measure different cognitive abilities. * A pre-registered study is a study in which context, aims, hypotheses and methodologies have been written down as an empirical paper, peer-reviewed and pre-accepted before research is undertaken. Pre-registrations are intended to reduce publication bias and reporting bias.

References:

Bond, A. B., Kamil, A. C., & Balda, R. P. (2007). Serial reversal learning and the evolution of behavioral flexibility in three species of north american corvids (Gymnorhinus cyanocephalus, Nucifraga columbiana,
Aphelocoma californica). Journal of Comparative Psychology, 121 (4), 372.
https://doi.org/10.1037/0735-7036.121.4.372

Coulon, A. (2019) Can context changes improve behavioral flexibility? Towards a better understanding of species adaptability to environmental changes. Peer Community in Ecology, 100019.
https://doi.org/10.24072/pci.ecology.100019

Logan, CJ, Lukas D, Bergeron L, Folsom M, & McCune, K. (2019). Is behavioral flexibility related to foraging and social behavior in a rapidly expanding species? In Principle Acceptance by PCI Ecology of the Version on 6 Aug 2019. http://corinalogan.com/Preregistrations/g_flexmanip.html

McCune KB, Blaisdell AP, Johnson-Ulrich Z, Lukas D, MacPherson M, Seitz BM, Sevchik A, Logan CJ (2023) Using repeatability of performance within and across contexts to validate measures of behavioral flexibility. EcoEvoRxiv, ver. 5 peer-reviewed and recommended by Peer Community in Ecology.
https://doi.org/10.32942/X2R59K

Reviews

Evaluation round #3

DOI or URL of the preprint: https://doi.org/10.32942/X2R59K
Version of the preprint: 3

Authors' reply, 24 May 2023

Hi Aurélie Coulon,

Thank you for the time and effort you have taken to help us improve this manuscript. I changed the citations throughout for the companion publication that stemmed from this research project (Logan et al. 2023), and updated the preprint at EcoEvoRxiv. I am excited for this piece to be recommended! Please let me know if there is anything else you need from me.

Best,

Kelsey (on behalf of all of the co-authors)

Decision by Aurélie Coulon, posted 19 May 2023, validated 21 May 2023

Updating of a reference

Dear Dr McCune and collaborators,

Maxime Dahirel and myself agree that you have fully addressed all the remaining concerns made on the previous version of your ms entitled "Using repeatability of performance within and across contexts to validate measures of behavioral flexibility". I am ready to publish my recommendation of this paper but before that, I wanted to give you the opportunity to update the reference to the other paper accompanying this one, which was recently recommended ("Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context").

Best,

Aurélie Coulon.

Reviewed by Maxime Dahirel, 16 May 2023

I thank and congratulate the authors for their new revision, which addresses all of my remaining concerns. I have nothing left to add, except reminding them that now that the other paper attached to this preregistration is recommended, they may want to update its reference (both year and details) in the final version of this one.

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.32942/X2R59K

Version of the preprint: 2

Authors' reply, 05 May 2023

Download author's reply

Download tracked changes file

Decision by Aurélie Coulon, posted 14 March 2023, validated 14 March 2023

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.

Reviewed by Maxime Dahirel, 16 February 2023

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 1

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.
- 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).

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.

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.

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.

MINOR COMMENTS

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.

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

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

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 March 2023

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.

Evaluation round #1

DOI or URL of the preprint: **https://doi.org/10.32942/osf.io/kevqp**
Version of the preprint: 1

Authors' reply, 01 February 2023

Round #1

by Aurélie Coulon, 14 Oct 2022 18:41
Manuscript: https://doi.org/10.32942/osf.io/kevqp

Revision needed for the preprint ”Repeatability of performance within and across contexts measuring behavioral flexibility”

Comment C1: Dear Dr McCune and collaborators,

Two reviewers have evaluated your preprint “Repeatability of performance within and across contexts measuring behavioral flexibility”. One of them is rather positive about your work. However, both of them made substantial comments on your ms, including methodological but also conceptual aspects. These remarks will be important to address. I also have a few minor requests, listed below.

Best,

Aurélie Coulon.
Recommender, PCI Ecology.

In the supplementary material, independent variables for P3b: I do not understand what IVs 3 & 4 are and so I don’t understand which variables were eventually included in the test of this hyp. Please clarify.

Response C1: We are sorry for the confusion here! The supplementary material consists of the portions of the original preregistration that pertain to the hypotheses addressed in this manuscript. Because we want the preregistration to remain unchanged after it has received In Principal Acceptance, we tried to instead point out subsequent deviations from the preregistration by writing a “Note” above each section. Here, the IVs that became redundant when we realized we needed to change the dependent variables are those listed as 3 (Latency to solve a new locus) and 4 (Number of trials to reverse a preference).

The variables used to test the P3b hypothesis were: Latency to switch a preference as a function of Reversal number (the number of times individuals switched a preference when the previously preferred color/locus was made non-functional) + Context (MAB plastic, MAB wood, reversal learning) + ID random effect. This section now reads:

“NOTE: Jul 2022 we changed the dependent variable to reflect the general latency to switch a preference (in any of the three tasks) and so IVs 3 (Latency to solve a new locus) & 4 (Number of trials to reverse a preference), below, are redundant. Furthermore, we did not include the touchscreen experiment in this manuscript (previously accounted for with IV 5, see the Deviations section). Therefore, despite being listed here in the preregistration as IVs we proposed to include in the P3b model, in our post-study manuscript we did not include these IVs in the final model. The IVs instead consisted of: Reversal (switch) number, Context (colored tubes, plastic multi-access box, wooden multi-access box) and ID (random effect because repeated measures on the same individuals).”

Comment C2: M&M, l.238 (and throughout the ms): to be consistent with the vocabulary used before, use “context” instead of “condition”.

Response C2: Good idea! There were three instances of “condition” throughout the post-study manuscript that we updated to “context”.

Comment C3: M&M: for test of H1, remind how many reversal trials the individuals were subjected to (i.e.

how many lines there are for each individual in the dataset).

Response C3: Good point. We added this text to Methods > Summary of Methods > Repeatability analysis:
"The reversal number for each grackle ranged between 6 to 11 (mean = 7.6) reversals, and the range was based on when individuals were able to pass two subsequent reversals in 50 or fewer trials, or (in 1 case) when we reached the maximum duration that we were permitted to keep grackles in the aviaries and they needed to be released."

Comment C4: Results: Hyp 1: it would be helpful for the reader to be able to visualize the distribution of values for the test of significance of the values of repeatability. Also, more information on the results of the test is needed (e.g. show variability within individuals, for each individual, like in Fig. 2).

Response C4: We added a new figure (Figure 2) to illustrate the results for Hypothesis 1. The plot shows the raw data on the number of trials to pass each reversal for each individual in a similar manner as the plot for Hypothesis 2. Then we added another plot (Figure 3) to show the histogram of results from the simulation test of significance of our repeatability value.

Comment C5: Hyp2: please add a summary table for the GLMM, showing variables are significant and what their estimates are.

Response C5: We are happy to add this table if you believe it is important. However, we ran the GLMM simply to estimate the variance components for performance within and between individuals, while accounting for potential confounds (the independent variables). This permits the calculation of the adjusted repeatability to determine whether individuals are consistent in their switching performance across contexts. Consequently, we considered the independent variable parameter estimates and significance values to be unrelated to the question we were testing.

Reviews
Reviewed by Maxime Dahirel, 13 Oct 2022 09:58

I have now read the manuscript entitled "Repeatability of performance within and across contexts measuring behavioral flexibility" by McCune et al. This manuscript stems from a three-way split of a previous large manuscript that has an in-principle preregistration acceptance.

In this context, the first and most important test is "does the present manuscript stands on its own, or should any insights in it be subsumed in one of the other manuscripts?". I believe this manuscript passes this test, by focusing on a tight, single question that is usefully treated as an separate short note.

Once this is out of the way, I have mostly one big methodological comment (COMMENT 3): while I agree the question asked is interesting and relevant, I don't think the methods used in the paper answer it, or at least not completely. The authors also ignore important methodological developments in the analysis of personality/repeated behavioral data. Note that it is possible that the analytical choices the authors made are valid and/or the only ones possible given limited available data, but it falls to them to make that explicit, especially in a sub-discipline where there is a de facto gold standard method to analyse such data (Dingemanse & Dochtermann, 2013; Dingemanse & Wright, 2020; O'Dea et al., 2022). Assuming the main qualitative results hold after revision, the Introduction and Discussion are already mostly clear, justified and to the point as is.

Please find below more details on this and other comments.

Comment D1: Please also note that many of the style and form comments I gave last month in my review of

the first manuscript out of this revision ("Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context") may also be relevant here, even if they are not copied below.

Response D1: We really appreciate you reviewing this work at multiple stages of this process. We revisited your comments from the previous version and list here some of the key changes we made to increase readability:

Added a Figure 1 showing the experimental devices

Formatted the article in a more traditional way with added Keywords and Methods sections in the usual places. Furthermore, we simplified the hypotheses and included them in-paragraph at the end of the intro rather than in a separate section.

—-

Comment D2: The way the authors write the manuscript, especially in the Abstract (line 25) and Introduction (line 60), it feels like temporal and/or contextual consistency is a sufficient condition to ensure the measured behaviours reflect latent intrinsic cognitive ability. It is not: it is a necessary but not sufficient condition. The only thing proven by a non-zero repeatability, is that behavioural expression is correlated across time/contexts. Even assuming that this correlation is due to an underlying intrinsic latent trait, this trait is not necessarily cognitive ability, and there is no reason to think there is only one latent trait involved anyway. The authors mention alternative causes a few times in the manuscript, so they are aware of this at least to an extent. However, this should be explicitly acknowledged in the Introduction, as "we show the thing we measure definitely reflects intrinsic cognition" and "we show the thing we measure is stable enough at an individual level that it's compatible with being intrinsic cognition" are not the same promises to the readers."

Response D2: This is a good point, we understand we cannot assume the repeatability in performance represents a cognitive trait. We updated the text as you suggested to be more cautious in these assertions (changes in bold):

Abstract:

"...**Although it can be difficult, or impossible, to definitively assign performance to one cognitive trait**, one way to validate that task structure **is more likely** to elicit performance based on the target cognitive trait is to assess the temporal and contextual repeatability of performance."

Introduction:

"**Cognitive traits are not directly observable and nearly all methods to quantify cognition use behavioral performance as a proxy for cognitive ability. Consequently, it is important to evaluate the validity of the chosen methods for quantifying a cognitive trait. To better understand whether performance on a type of task is likely to reflect a target cognitive trait (i.e., that the method has construct validity), researchers can test for repeatability in individual performance within and across tasks [@volter2018comparative]. However,** while many cognitive abilities have been tested, and various methods used, it is rare for one study to repeatedly test individuals with the same method or use multiple methods to test for a given cognitive ability. **Using only one method to measure a cognitive ability could be problematic because it is hard to discern whether** non-target cognitive, personality, or motivational **factors may be the cause of variation in** performance on the task [@morand2016studying]. For example, the success of pheasants on multiple similar and different problem-solving tasks was related to individual variation in persistence and motivation, rather than problem solving ability [@van2016problem]. Additionally, performance on cognitive tasks can be affected by different learning styles, where individuals can vary in their perception of the salience of stimuli within a task, the impact of a reward (or non-reward) on future behavior, or the propensity to sample alternative stimuli [@rowe2014measuring]. **By assessing the temporal and contextual repeatability of performance researchers can quantify the proportion of variation in performance**

that is attributable to consistent individual differences likely to reflect the level of the cognitive trait relative to other ephemeral factors that affect individual performance [@cauchoix2018repeatability]."

Comment D3: Introduction, especially the first paragraph: here a good part of the argument relies on the fact that cognition, personality, motivation, learning ability... are separate entities/concepts. Putting aside the fact the current study may not be able to separate these (see comment 1 above), I think I am OK with that as a premise. However, I, and probably other readers, would definitely need a definition of cognition, and possibly some of the other terms, to be sure I am OK with that.

Response D3: We added a definition for cognition as follows:

"**As a result, we have come to understand cognition as the process of acquiring information, followed by storage, retrieval, and use of that information for guiding behavior [@shettleworth2010cognition].** Consequently, evidence now exists that various species possess cognitive abilities in both the physical [e.g. object permanence: @salwiczek2009development; causal understanding: @taylor2012end] and social domains [e.g. social learning: @hoppitt2012identification; transitive inference: @maclean2008social]."

Furthermore, we elaborated on our definition of behavioral flexibility in paragraph 3 of the introduction:

"Behavioral flexibility, the ability to change behavior when circumstances change, is a general cognitive ability that likely affects interactions with both the social and physical environment [@bond2007serial]. **Although by definition behavioral flexibility incorporates plasticity in behavior through learning, there is also evidence that the ability to change behavior could be an inherent trait that varies among individuals and species. For example, the pinyon jay - a highly social species of corvid - made fewer errors in a serial reversal learning task than the more asocial Clark's nutcracker or Woodhouse's scrub-jay, but all three species exhibited similar learning curves over successive reversals [@bond2007serial]. This indicates that the three species differed in the level of the inherent ability, but were similar in plasticity of performance through learning.**"

Comment D4: I am not convinced the methods described in the manuscript can answer the question asked for several reasons. Before I go in quite a lot of details about this, I need to preface that while these comments reflect recommended practices in the subfield of let's say "individual differences in repeated behaviours and other labile traits", alternatives may be valid in certain contexts, and that may include the analytical choices the authors made. However, when such strong and principled standard methods exist, authors must carefully consider them before deciding on alternatives:

3a – As mentioned comment 4 of my review of manuscript 1, the authors do not account for individual random slopes when they should (unless data are insufficient). Copying that comment verbatim here:

"Regarding your response to original review comment D9: I was happy to see that you say that including a individual-level random slope effect of reversal number does not change the conclusion. However, based on the text (line 297) and the code (https://github.com/corinalogan/grackles/blob/5b220d237542053bfa51673ec38116497affe55c/Files/Preregistrations/g_flexmanip.Rmd#L219), I actually see no evidence that you actually did fit the model with random slopes. See https://stat.ethz.ch/pipermail/r-sig-mixed-models/2013q1/019896.html for an example of the expected syntax for the inclusion of individual-level random slopes in MCMCglmm, and please refer back to the citations I mentioned in the original comment, for the importance of correctly accounting for this (and also Harrison et al 2018, https://peerj.com/articles/4794, chapter "Choosing random effects II: random slopes" for a review). Apologies in advance if I missed the correct model somehow.

Note that I acknowledge that in some cases the dataset may be too small to estimate these random slopes accurately (see the aforementioned Harrison et al. reference, but then, the better option is to acknowledge it explicitly, rather than ignoring it."

The authors are actually aware that failing to account for this temporal dependency may affect their conclusion based on repeatability (see lines 246-247); they failed to realise that this (random slopes) would be a principled way to account for it. They are also aware of the context of behavioural reaction norms: random slopes are the tool to account for individual variation in behavioural reaction norms, the "environment" here being "experience" (see fig 1 for instance in Schielzeth & Nakagawa, 2022).

Of course, adding random slopes complicates the estimation of repeatability from model variance components, since it means repeatability may vary through time/reversal number: some individuals may be better at reversing within a reversal trial on average (intercept random effect, if reversal number is centred), some individuals may be better at learning faster from one reversal to the next ("at learning to learn", in a way: slope random effect), and these two may or may not be positively correlated. There is a substantial amount of literature on how to treat that (Brommer, 2013; Johnson, 2014; O'Dea et al., 2022; Schielzeth & Nakagawa, 2022). Note that some approaches to variation partitioning may be quite involved and overly complex, and keep in mind what is your question and what is the variance component you are the most interested in: for instance, assuming you are only interested in the intercept random effect, the simplest, but potentially good enough for your purpose, would be to consider (Johnson, 2014) and estimate repeatability at the average time/reversal number.

Response D4: Thank you for your detailed suggestions and explanations on this point. We read many of the papers you suggested here, and below, to bring ourselves up-to-speed on these techniques.

In terms of temporal dependency, we included the reversal number as an independent variable (and not a random effect) in our model because we are interested in controlling for the temporal dependency rather than estimating individual differences in "learning to learn". Given that our sample size is only 9 individuals, we wanted to focus on the average repeatability of performance across reversals. We hope to build on this in later studies to assess consistent individual differences in rate of learning across tasks, but this would require more subjects and longer testing periods (i.e. more reversals in different kinds of tasks) than we had the ability to prioritize in this study.

We did attempt to run the model with a random slope included, however we could not get the model to converge with our sample size and the uninformative priors that were preregistered. We felt most comfortable using the preregistered methods to avoid biasing the model output.

We added this information to the section Methods > Summary of Methods > Repeatability analysis:

"By design in the serial reversal learning experiment, to reach the experiment ending criteria grackles became faster at switching across serial reversals. **We did attempt to run a model that additionally included a random slope to test whether there were consistent individual differences in the rate that grackles switched their preferences across reversals. However, we could not get the model to converge with our sample size and the uninformative priors that were preregistered. We felt most comfortable using the preregistered methods to avoid biasing the model output.** Instead, to determine the statistical significance of the repeatability value, while accounting for this non-independence of a change in reversal speed over time, we compared the actual performance on the number of trials to switch a preference in each reversal to simulated data where birds performed randomly within each reversal."

Comment D5: I am unclear about why the authors try to combine two different data types (from colour and MAB tests) into one single model, having to distort one data type (converting counts to times) in the process.

As mentioned in several of the references above (Dingemanse & Dochtermann, 2013; Dingemanse & Wright, 2020; O'Dea et al., 2022), the principled approach would be to fit a multivariate mixed model where each repeated behaviour has its own submodel, which may be different families. This approach also allows separate answers to the questions "are color reversal and MAB tests repeatable?" and "are the repeatable components of these traits correlated in a way that suggests an underlying common latent variable?", as it should be. See also (Martin et al., 2019) but also and especially (Bailey et al., 2021) and references therein for potential avenues to test this.

Response D5: When comparing performance across tests, we thought it was not unusual to convert the measured variable/criteria to the same scale for all tasks (e.g. Amy et al. 2012; Griffin et al. 2013). For both the MAB and reversal learning tasks, performance in the form of number of trials and latency was available and we chose latency because as a continuous (rather than discrete, like trial number) variable it contained more information.

Thank you for pointing us in the direction of this statistical approach. However, given our small sample size, we do not think it is appropriate for our data. Dingemanse & Wright (2020) specifically say: "Unfortunately, multivariate mixed-effects models are extremely data hungry, implying that they can only be applied to large data sets". Which, as you point out in Comment D7 below, is not the case for our data.

Comment D6: line 227, the authors mention that they used a Gaussian model on log-transformed data rather than the preregistered, and canonical, Poisson model. First, if that's the only remedy possible to the identified problem, this is OK. However, I believe it is not. To account for overdispersion, the authors may first reach for solutions such as observation-level random effects or negative binomial models (Harrison, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2010). Second, one must keep in mind that strictly speaking, heteroskedasticity is a feature, not a bug, of count distributions like Poisson (where variance = mean) or negative binomial. The question is whether there is more or less heteroscedasticity than expected for the distribution. Packages such as DHARMa provide user-friendly tests for that. Finally, and possibly most importantly, any overdispersion or heteroskedasticity or similar issues may be because the model is incomplete because it ignores temporal non-independence, for instance and may vanish once random slopes (see 3a) are properly included.

Response D6: We tested for model fit using the DHARMa package and it indicated significant heteroscedasticity and overdispersion for a Poisson model. However, there are no significant problems in model fit detected by the DHARMa package when we use a gaussian distribution and log-transformed dependent variable. Visualizing the trace of the parameters in the actual model also confirms that the gaussian distribution and log-transformed dependent variable produce a good model fit.

At your suggestion, we attempted to use observation-level random effects (OLRE) to control the overdispersion. However, we first had model fit issues where we were required to set strong priors to get the model to run. Secondly, the OLRE now accounts for all residual variance and we do not see how it is then possible to calculate repeatability (grouping factor/grouping factor + residual variance). If we use OLRE in the calculation so that the repeatability denominator still represents total variance, the repeatability value is significantly larger and the credible interval is so large as to be uninformative (R = 0.5, CI = 0.008-1). We sought advice from you originally on how to calculate repeatability from MCMCglmm output, so perhaps you can give some guidance on this?

Finally, when reading some of your suggested literature it seems that, per Schielzeth et al. 2020, violation of

model assumptions has a much more minor impact on accuracy and precision of results in Gaussian compared to Poisson distributions. Additionally, even though our response variable is a count, it does not fit the usual assumptions of the Poisson distribution because the number of trials to pass a reversal is bounded on the lower end of the scale at 20, and for later reversals it is also bounded on the upper end by 50. According to McElreath 2015, there is no loss of information when Poisson data are log-transformed (pg. 239). Consequently, it seems like the overall better option is to log-transform our data so they fit a Gaussian distribution rather than include OLRE in a Poisson model.

MINOR COMMENTS:

Comment D7: As in the previous manuscript, a (small) discussion of how the interpretation of the results is affected by the low number of individuals (especially how it leads to uncertainty and wide CIs around parameters of interest) would be welcome

Response D7: We added explicit mention of our small sample size as a study limitation in the discussion:

In the second paragraph:

"**Funding and logistical limitations result in few researchers assessing** the appropriateness of their methods by testing construct validity through convergent (similar performance across similar tasks) and discriminant validity (different performance across different tasks). **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."

As well as in the third paragraph:

"That behavioral flexibility is both repeatable within individuals across reversals, indicating it is an inherent trait, as well as being manipulatable through serial reversals, aligns with the idea of behavioral reaction norms [@sih2013understanding]. This idea states that individuals can show consistent individual differences in the baseline or average values of a trait of interest across time or contexts, but the plasticity in the expression of the trait can also consistently vary among individuals. **Due to our small sample size, we were not able to explicitly test for behavioral reaction norms, but this is an important next step in understanding consistent individual variation in behavioral flexibility in relation to rapid environmental change.**"

Comment D8: it might be best to precise what type of models and software have been used in the "main" text, rather than leave it to the "supplement" that details the pre-registration original text. The same may be true for some details of the experimental protocol.

Response D8: We see how this would be helpful. We added more methodological information per your comment D12, below, and we also added analysis methods in METHODS > Summary of Methods > Repeatability Analysis

"We used the DHARMa package [@hartig2019dharma] in R to test whether our model fit our data and was not heteroscedastic, zero-inflated or over-dispersed. We used the MCMCglmm package [@hadfield2010mcmc], with uninformative priors, to model the relationships of interest for our two hypotheses."

Comment D9: Line 80: the authors mention 2 hypotheses, but there is only one described in the paragraph (in the sense that there a "first", but never a "second" or "then". From context one might surmise the authors treat temporal repeatability and contextual repeatability as 2 hypotheses. I may not agree, but if that is what

the authors imply, then the paragraph should be rewritten to make structurally clear.)

Response D9: Apologies for the confusion. We clarified that we are referring to two hypotheses here (changes in bold):

"We tested two hypotheses about the construct validity of the reversal learning method as a measure of behavioral flexibility in the great-tailed grackle (*Quiscalus mexicanus*; hereafter "grackle"). First, we determined whether performance on a color reversal learning task represents an inherent trait by assessing the repeatability of performance across serial reversals (temporal repeatability). **Second, we determined whether the inherent trait measured by color reversal learning is likely to represent behavioral flexibility by assessing the cross-contextual repeatability of performance on this task with another task also thought to measure flexibility.** Our previous research found that behavioral flexibility does affect innovation ability on a multi-access box [@logan2022flexmanip], so **here our second hypothesis tested whether individuals show** contextual repeatability of flexibility by comparing performance on the reversal learning task to the latency of solution switching on two different multi-access boxes (Fig. 1)."

Comment D10: Line 80: please put scientific names in italic

Response D10: Thank you for pointing out this oversight! It is now fixed.

Comment D11: Line 124-125: please put spaces between numbers and units

Response D11: We added spaces as suggested.

Comment D12: Line 131: please clarify that the food stayed in the same color until the bird learned the association, then stayed on the other color for reversal. This is obvious when one stops for a few seconds to think about it, sure, but the way the text is currently written makes it more implicit than it should.

Response D12: We added the following text to clarify this point:

"Briefly, we trained grackles to search in one of two differently colored containers for food (Fig. 1a). **We used a random number generator to select the color (e.g. light gray) of the container that would consistently contain a food reward across the initial trials. Within each trial, grackles could choose only one container to look in for food. Eventually,** grackles showed a significant preference for **the rewarded** color **container** (minimum of 17 out of 20 correct choices), **completing the initial discrimination trials. We then** switched the location of the food to the container of the other color (a reversal). **The food reward was then consistently located in the container of this second color (e.g. dark gray) across trials until the grackles learned to switch their preference, after which we would again reverse the food to the original colored container (e.g. light gray) and so on back and forth until they passed the serial reversal learning criterion.**"

Comment D13: Fig 2 and Table 1. Simpler, more standard figure captions may be welcome. E.g. for Table 1 "Table 1 - Repeatability of learning performance across time within and across contexts"

Response D13: Note that Fig 2 is now Fig 4 (per a comment from the Recomender, above) and we removed Table 1 per your Comment D14, below. We modified our figure caption as you suggested:

"Figure 4: **Grackle performance on the different contexts for measuring behavioral flexibility:** multi-access box (MAB) plastic (square symbol), MAB wood (triangle symbol), and reversal learning with color tubes

(star symbol)..."

Comment D14: Table 1: depending on the analyses after revision, table 1 may or may not be necessary. In the current manuscript, most of that information is already given in text. If CIs were just added in-text along with mean repeatabilities, there would be no need for table 1. (as an aside, please precise the confidence interval width? Is it the standard 95% or another one, as in some analyses in the previous manuscript)

Response D14: Great point! We removed Table 1 because it is relatively uninformative. We added CIs to the text in the Results section, and we specified they were 95% credible intervals.

Comment D15: Figure 2: The axis is not in seconds, since there are negative values. I assume it's actually log(seconds). Please correct the legend of the axis or backtransform the data before plotting

Response D15: Sorry for this error. We changed the y-axis of the figure to represent the form of the data (centered and scaled seconds). Note that this is now Figure 3.

REFERENCES

Bailey, J. D., King, A. J., Codling, E. A., Short, A. M., Johns, G. I., & Fürtbauer, I. (2021). "Micropersonality" traits and their implications for behavioral and movement ecology research. Ecology and Evolution, 11(7), 3264–3273. https://doi.org/10.1002/ece3.7275

Brommer, J. E. (2013). Variation in plasticity of personality traits implies that the ranking of personality measures changes between environmental contexts: Calculating the cross-environmental correlation. Behavioral Ecology and Sociobiology, 67(10), Article 10. https://doi.org/10.1007/s00265-013-1603-9

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. Journal of Animal Ecology, 82(1), Article 1. https://doi.org/10.1111/1365-2656.12013

Dingemanse, N. J., & Wright, J. (2020). Criteria for acceptable studies of animal personality and behavioural syndromes. Ethology, 126(9), Article 9. https://doi.org/10.1111/eth.13082

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ, 2, e616. https://doi.org/10.7717/peerj.616

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods in Ecology and Evolution, 5(9), Article 9. https://doi.org/10.1111/2041-210X.12225

Martin, J. S., Massen, J. J. M., Šlipogor, V., Bugnyar, T., Jaeggi, A. V., & Koski, S. E. (2019). The EGA+GNM framework: An integrative approach to modelling behavioural syndromes. Methods in Ecology and Evolution, 10(2), Article 2. https://doi.org/10.1111/2041-210X.13100

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. Journal of The Royal Society Interface, 14(134), Article 134. https://doi.org/10.1098/rsif.2017.0213

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

O'Dea, R. E., Noble, D. W. A., & Nakagawa, S. (2022). Unifying individual differences in personality, predictability and plasticity: A practical guide. Methods in Ecology and Evolution, n/a(n/a). https://doi.org/10.1111/2041-210X.13755

Schielzeth, H., & Nakagawa, S. (2022). Conditional repeatability and the variance explained by reaction norm variation in random slope models. Methods in Ecology and Evolution, 13(6), 1214–1223. https://doi.org/10.1111/2041-210X.13856

Reviewed by Aparajitha Ramesh, 07 Oct 2022 14:19

Report on "Repeatability of performance within and across contexts measuring behavioral flexibility" by McCune et al., 2022

Comment R1: The study aims to test whether behavioral flexibility in terms of reversal learning is repeatable within a general context of flexibility of behavior. In addition, they also motivate their current study with the fact that construct validity, which is whether the trait of interest is actually being tested is captured by the experimental methodology. I very much enjoyed reading the pre-registration and the current paper and commend the authors on their endeavour towards open science. Furthermore, I very much liked the explanation on how the present study is integrated into a larger study and the details including codes for analyses are also made transparent. Overall I think this paper is an excellent example of good practises in science and also addresses an interesting question.

The results stemming out of this study are also interesting for research lines on animal personality. Although my impression of the manuscript is quite positive, I have two major concerns namely (1) The focus of the study (validating methodology or testing the biological phenomenon of behavioural flexibility) (2) Repeatability estimates for temporal repeatability (low and overlaps with zero). Other than these, I have a few minor suggestions which may be useful.

Response R1: Thank you for your kind words! We addressed your concerns, below, and we think the manuscript is now much improved.

Comment R2: The focus of the current study: At several points in the introduction, emphasis is made on the methodology used for testing behavioral flexibility and testing construct validity to ensure the method is testing flexibility regardless of state, personality etc. (Eg. On line 60). However, I also see that the authors have expectations on the repeatability of behavioral flexibility as a biological phenomenon. It is therefore confusing to me how the same data is used to address both questions and whether it can delineate the two. For example, if we find that behavioral flexibility are indeed repeatable over time (significant proportion of within individual variation), does temporal repeatability give an account of consistent variation in inherent flexibility and learning in animals or does it validate the methodology used for testing? I do believe that a philosophical discussion on this is useful and methodological validation can be obtained also by relating to the behaviors in the wild. But I would appreciate if there is one clear line of thoughts here.

Response R2: You make an excellent point. We see that we were not specific enough in the introduction on this issue, which led to some confusion. The focus of the study is on validating methods currently popular for testing the cognitive ability we call behavioral flexibility. Determining the construct validity of a method can be accomplished, in part, through assessing the repeatability of performance across time and across different contexts thought to measure behavioral flexibility.

First, evaluating the repeatability of performance across time informs whether the performance is consistent within individuals and therefore unlikely to be due to ephemeral motivational or environmental influences. Secondly, evaluating the repeatability of performance across context informs whether the inherent trait governing performance is likely to be the trait we call behavioral flexibility. In other words, both repeatability across time and context are needed to establish that a method has construct validity, but the first measure determines whether the method measures an inherent trait, and the second measure assesses the likelihood that the measured trait is in fact the trait that the method is hypothesized to measure.

To remedy the confusion we modified the wording and flow in the second paragraph of the introduction (additions in bold face font):

"**Cognitive traits are not directly observable and nearly all methods to quantify cognition use behavioral performance as a proxy for cognitive ability. Consequently, it is important to evaluate the validity of the chosen methods for quantifying a cognitive trait. To better understand whether performance on a type of task is likely to reflect a target cognitive trait (i.e., that the method has construct validity), researchers can test for repeatability in individual performance within and across tasks [@volter2018comparative]. However**, while many cognitive abilities have been tested, and various methods used, it is rare for one study to repeatedly test individuals with the same method or use multiple methods to test for a given cognitive ability. **Using only one method to measure a cognitive ability could be problematic because it is hard to discern whether** non-target cognitive, personality, or motivational **factors may be the cause of variation in** performance on the task [@morand2016studying]. For example, the success of pheasants on multiple similar and different problem-solving tasks was related to individual variation in persistence and motivation, rather than problem solving ability [@van2016problem]. Additionally, performance on cognitive tasks can be affected by different learning styles, where individuals can vary in their perception of the salience of stimuli within a task, the impact of a reward (or non-reward) on future behavior, or the propensity to sample alternative stimuli [@rowe2014measuring]. **By assessing the temporal and contextual repeatability of performance researchers can quantify the proportion of variation in performance that is attributable to consistent individual differences likely to reflect the level of the cognitive trait relative to other ephemeral factors that affect individual performance [@cauchoix2018repeatability].**"

As well as adding more specific wording to our third paragraph of the introduction where we describe our hypotheses for this manuscript:

"We tested two hypotheses about the **construct** validity of the reversal learning method as a measure of behavioral flexibility in the great-tailed grackle (***Quiscalus mexicanus***; hereafter "grackle"). First, we determined whether performance on a **color** reversal learning task represents an inherent trait by assessing the repeatability of performance across serial reversals (temporal repeatability). **Secondly, we determined whether the inherent trait measured by color reversal learning is likely to represent behavioral flexibility by assessing the cross-contextual repeatability of performance on this task with another task also thought to measure flexibility.** Our previous research found that behavioral flexibility does affect innovation ability on a multi-access box [@logan2022flexmanip], **so here our second hypothesis tested whether individuals show** contextual repeatability of flexibility by comparing performance on the color reversal learning task to the latency of solution switching on two different multi-access boxes (Fig. 1)."

Comment R3: Temporal repeatability in Table 1: Temporal repeatability is quite low. Moreover, the credible intervals overlaps with 0, which one would interpret as not significantly different from 0. So I was wondering how p.value shows that it is significant.

In line 247, it is explained that the significance of repeatability was calculated by comparing to simulated data where birds performed randomly within each reversal – Is this the same as testing when confidence intervals for R is calculated by bootstrapping and statistical significance is tested against the null hypothesis that R=0 (as done in rptR package, Stoffel et al. 2017)?

Additionally, the p.value in the main text (line 250, p=0.001) is also different from the p.value denoted in the table 1 (p=0.01).

Overall, it may be important to verify whether the repeatability value is large enough and if it is different from zero to determine whether P3a alternative should be worked out or alternative explanations are needed.

Response R3: The actual number for our lower credible interval (CI) was 4.64 x 10-16, but we rounded it to zero for brevity. Repeatability is a ratio, so the CI will never reach below zero, but it is common for the lower end of the CI to approach zero (Bell et al. 2009). That is why it is important to determine the statistical significance of the repeatability value through permutation tests.

A meta-analysis of repeatability of cognitive performance from 5 years ago found that R can range from 0.15 - 0.28 (Cauchoix et al. 2018). Our R value for temporal repeatability falls just outside of that range. So we used the permutation test to determine whether R=0.13 was too low to be significantly repeatable. This method is very similar to the permutation test conducted in the rptR package for estimating the significance of the repeatability estimate. However, we could not use that (more simple) method because we had to account for the non-random aspect of our experimental design where we set the criteria for ending the serial reversal learning experiment as when a grackle passed two sequential reversals in 50 or fewer trials.

Thank you for pointing out the mis-match in p-values between the table and the text. That appears to be a typo and it is now fixed in the text - but note that we removed Table 1 in response to a comment from another reviewer. CIs are now placed in the text of the results section with the reporting of the R values.

Comment R4: Title should be changed to address the focus (see comment (1)). If this is methodological, it would be nice to explicitly have it in the title for greater comprehension

Response R4: We modified the title to emphasize the methodological framing of this manuscript. It now reads:

"Using repeatability of performance within and across contexts **to validate** measures **of** behavioral flexibility"

Comment R5: I am not familiar with this but maybe keywords are missing?

Response R5: We added the following keywords after the Abstract:

"Behavioral flexibility, repeatability, construct validity, animal cognition"

Comment R6: Line 29: Add scientific name

Response R6: Thank you for pointing out this oversight! It is now fixed.

Comment R7: Line 38: Maybe a personal preference, but perhaps the starting of the paragraph can be improved and sentence can be made shorter.

Response R7: You are right, this is an overly complicated and wordy sentence. We separated it into discrete sentences such that the first part of this paragraph now reads:

"Research on the cognitive abilities of non-human animals is important for several reasons. By understanding animal cognitive abilities, we can clarify factors that influenced the evolution of human cognition, the mechanisms that relate cognition to ecological and evolutionary dynamics, or we can use the knowledge to facilitate more humane treatment of captive individuals [@shettleworth2010cognition]."

Comment R8: Line 80: Scientific name should be italicised

Response R8: Thank you for pointing out this oversight! It is now fixed.

Comment R9: Line 80: 'First' is not followed by 'Second' or 'Then' – maybe remove it if the hypotheses are not discussed here but later.

Response R9: Apologies for the confusion. We have clarified that we are referring to two hypotheses here (changes in bold):

"...**here our second hypothesis tested whether individuals show** contextual repeatability of flexibility by comparing performance on the reversal learning task to the latency of solution switching on two different multi-access boxes (Fig. 1)."

Comment R10: Line 117: Secondly should be preceded by 'Firstly'. Here, the Secondly describes the hypothesis mentioned in the introduction. Could the two hypotheses that are being dealt with in this paper be made clearer? Also in Line 80 (see previous comment)

Response R10: Thank you for pointing this out. This line now reads:
"**Our first hypothesis considered whether** behavioral flexibility (as measured by reversal learning of a color preference) would be repeatable within individuals across serial reversals."

Reference:
Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods in Ecology and Evolution, 8(11), 1639-1644.
Download tracked changes file

Decision by Aurélie Coulon, posted 14 October 2022, validated 14 October 2022

Revision needed for the preprint "Repeatability of performance within and across contexts measuring behavioral flexibility"

Dear Dr McCune and collaborators, Two reviewers have evaluated your preprint "Repeatability of performance within and across contexts measuring behavioral flexibility". One of them is rather positive about your work. However, both of them made substantial comments on your ms, including methodological but also conceptual aspects. These remarks will be important to address. I also have a few minor requests, listed below. Best, Aurélie Coulon.
Recommender, PCI Ecology.
In the supplementary material, independent variables for P3b: I do not understand what IVs 3 & 4 are and so I don't understand which variables were eventually included in the test of this hyp. Please clarify. M&M, l.238 (and throughout the ms): to be consistent with the vocabulary used before, use "context" instead of "condition". M&M: for test of H1, remind how many reversal trials the individuals were subjected to (i.e. how many lines there are for each individual in the dataset). Results:
Hyp 1: it would be helpful for the reader to be able to visualize the distribution of values for the test of significance of the values of repeatability. Also, more information on the results of the test is needed (e.g. show variability within individuals, for each individual, like in Fig. 2).
Hyp2: please add a summary table for the GLMM, showing variables are significant and what their estimates are.

Reviewed by Maxime Dahirel, 13 October 2022

I have now read the manuscript entitled "Repeatability of performance within and across contexts measuring behavioral flexibility" by McCune et al. This manuscript stems from a three-way split of a previous large manuscript that has an in-principle preregistration acceptance.

In this context, the first and most important test is "does the present manuscript stands on its own, or should any insights in it be subsumed in one of the other manuscripts?". I believe this manuscript passes this test, by focusing on a tight, single question that is usefully treated as an separate short note.

Once this is out of the way, I have mostly one big methodological comment (COMMENT 3): while I agree the question asked is interesting and relevant, I don't think the methods used in the paper answer it, or at least not completely. The authors also ignore important methodological developments in the analysis of personality/repeated behavioral data. Note that it is possible that the analytical choices the authors made are valid and/or the only ones possible given limited available data, but it falls to them to make that explicit, especially in a sub-discipline where there is a de facto gold standard method to analyse such data (Dingemanse & Dochtermann, 2013; Dingemanse & Wright, 2020; O'Dea et al., 2022). Assuming the main qualitative results hold after revision, the Introduction and Discussion are already *mostly* clear, justified and to the point as is.

Please find below more details on this and other comments. Please also note that many of the style and form comments I gave last month in my review of the first manuscript out of this revision ("Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context") may also be relevant here, even if they are not copied below.

—-

COMMENT 1

The way the authors write the manuscript, especially in the Abstract (line 25) and Introduction (line 60), it feels like temporal and/or contextual consistency is a sufficient condition to ensure the measured behaviours reflect latent intrinsic cognitive ability. It is not: it is a **necessary but not sufficient** condition. The only thing proven by a non-zero repeatability, is that behavioural expression is correlated across time/contexts. Even assuming that this correlation is due to an underlying intrinsic latent trait, this trait is not necessarily cognitive ability, and there is no reason to think there is only one latent trait involved anyway. The authors mention alternative causes a few times in the manuscript, so they are aware of this at least to an extent. However, this should be explicitly acknowledged in the Introduction, as "we show the thing we measure definitely reflects intrinsic cognition" and "we show the thing we measure is stable enough at an individual level that it's compatible with being intrinsic cognition" are not the same promises to the readers.

COMMENT 2

Introduction, especially the first paragraph: here a good part of the argument relies on the fact that cognition, personality, motivation, learning ability… are separate entities/concepts. Putting aside the fact the current study may not be able to separate these (see comment 1 above), I think I am OK with that as a premise. However, I, and probably other readers, would definitely need a definition of cognition, and possibly some of the other terms, to be sure I am OK with that.

COMMENT 3

I am not convinced the methods described in the manuscript can answer the question asked for several reasons. Before I go in quite a lot of details about this, I need to preface that while these comments reflect recommended practices in the subfield of let's say "individual differences in repeated behaviours and other labile traits", alternatives may be valid in certain contexts, and that may include the analytical choices the authors made. However, when such strong and principled standard methods exist, authors must carefully consider them before deciding on alternatives:

3a – As mentioned comment 4 of my review of manuscript 1, the authors do not account for individual random slopes when they should (unless data are insufficient). Copying that comment verbatim here:

"Regarding your response to original review comment D9: I was happy to see that you say that including

a individual-level random slope effect of reversal number does not change the conclusion. However, based on the text (line 297) and the code (https://github.com/corinalogan/grackles/blob/5b220d237542053bfa51673ec38116497affe55c/Files/Preregistrations/g_flexmanip.Rmd#L219), I actually see no evidence that you actually did fit the model with random slopes. See https://stat.ethz.ch/pipermail/r-sig-mixed-models/2013q1/019896.html for an example of the expected syntax for the inclusion of individual-level random slopes in MCMCglmm, and please refer back to the citations I mentioned in the original comment, for the importance of correctly accounting for this (and also Harrison et al 2018, https://peerj.com/articles/4794, chapter "Choosing random effects II: random slopes" for a review). Apologies in advance if I missed the correct model somehow.

Note that I acknowledge that in some cases the dataset may be too small to estimate these random slopes accurately (see the aforementioned Harrison et al. reference, but then, the better option is to acknowledge it explicitly, rather than ignoring it."

The authors are actually aware that failing to account for this temporal dependency may affect their conclusion based on repeatability (see lines 246-247); they failed to realise that this (random slopes) would be a principled way to account for it. They are also aware of the context of behavioural reaction norms: random slopes are the tool to account for individual variation in behavioural reaction norms, the "environment" here being "experience" (see fig 1 for instance in Schielzeth & Nakagawa, 2022).

Of course, adding random slopes complicates the estimation of repeatability from model variance components, since it means repeatability may vary through time/reversal number: some individuals may be better at reversing within a reversal trial on average (intercept random effect, if reversal number is centred), some individuals may be better at learning faster from one reversal to the next ("at learning to learn", in a way: slope random effect), and these two may or may not be positively correlated. There is a substantial amount of literature on how to treat that (Brommer, 2013; Johnson, 2014; O'Dea et al., 2022; Schielzeth & Nakagawa, 2022). Note that some approaches to variation partitioning may be quite involved and overly complex, and keep in mind what is your question and what is the variance component you are the most interested in: for instance, assuming you are only interested in the intercept random effect, the simplest, but potentially good enough for your purpose, would be to consider (Johnson, 2014) and estimate repeatability at the average time/reversal number.

3b – I am unclear about why the authors try to combine two different data types (from colour and MAB tests) into one single model, having to distort one data type (converting counts to times) in the process.

As mentioned in several of the references above (Dingemanse & Dochtermann, 2013; Dingemanse & Wright, 2020; O'Dea et al., 2022), the principled approach would be to fit a multivariate mixed model where each repeated behaviour has its own submodel, which may be different families. This approach also allows separate answers to the questions "are color reversal and MAB tests repeatable?" and "are the repeatable components of these traits correlated in a way that suggests an underlying common latent variable?", as it should be. See also (Martin et al., 2019) but also and especially (Bailey et al., 2021) and references therein for potential avenues to test this.

3c – line 227, the authors mention that they used a Gaussian model on log-transformed data rather than the pregistered, and canonical, Poisson model. First, if that's the only remedy possible to the identified problem, this is OK. However, I believe it is not. To account for overdispersion, the authors may first reach for solutions such as observation-level random effects or negative binomial models (Harrison, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2010). Second, one must keep in mind that strictly speaking, heteroskedasticity is a feature, not a bug, of count distributions like Poisson (where variance = mean) or negative binomial. The question is whether there is more or less heteroscedasticity than expected for the distribution. Packages such as DHARMa provide user-friendly tests for that. Finally, and possibly most importantly, any overdispersion or heteroskedasticity or similar issues may be *because the model is incomplete because it ignores temporal non-independence, for instance* and may vanish once random slopes (see 3a) are properly included.

MINOR COMMENTS:

COMMENT 4:

As in the previous manuscript, a (small) discussion of how the interpretation of the results is affected by the low number of individuals (especially how it leads to uncertainty and wide CIs around parameters of interest) would be welcome

COMMENT 5:

it might be best to precise what type of models and software have been used in the "main" text, rather than leave it to the "supplement" that details the pre-registration original text. The same may be true for some details of the experimental protocol.

COMMENT 6:

Line 80: the authors mention 2 hypotheses, but there is only one described in the paragraph (in the sense that there a "first", but never a "second" or "then". From context one might surmise the authors treat temporal repeatability and contextual repeatability as 2 hypotheses. I may not agree, but if that is what the authors imply, then the paragraph should be rewritten to make structurally clear.)

COMMENT 7:

Line 80: please put scientific names in italic

COMMENT 8:

Line 124-125: please put spaces between numbers and units

COMMENT 9:

Line 131: please clarify that the food stayed in the same color until the bird learned the association, then stayed on the other color for reversal. This is obvious when one stops for a few seconds to think about it, sure, but the way the text is currently written makes it more implicit than it should.

COMMENT 10:

Fig 2 and Table 1. Simpler, more standard figure captions may be welcome. E.g. for Table 1 "Table 1 - Repeatability of learning performance across time within and across contexts"

COMMENT 11:

Table 1: depending on the analyses after revision, table 1 may or may not be necessary. In the current manuscript, most of that information is already given in text. If CIs were just added in-text along with mean repeatabilities, there would be no need for table 1. (as an aside, please precise the confidence interval width? Is it the standard 95% or another one, as in some analyses in the previous manuscript)

COMMENT 12:

Figure 2: The axis is not in seconds, since there are negative values. I assume it's actually log(seconds). Please correct the legend of the axis or backtransform the data before plotting

REFERENCES

Bailey, J. D., King, A. J., Codling, E. A., Short, A. M., Johns, G. I., & Fürtbauer, I. (2021). "Micropersonality" traits and their implications for behavioral and movement ecology research. Ecology and Evolution, 11(7), 3264–3273. https://doi.org/10.1002/ece3.7275

Brommer, J. E. (2013). Variation in plasticity of personality traits implies that the ranking of personality measures changes between environmental contexts: Calculating the cross-environmental correlation. Behavioral Ecology and Sociobiology, 67(10), Article 10. https://doi.org/10.1007/s00265-013-1603-9

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. Journal of Animal Ecology, 82(1), Article 1. https://doi.org/10.1111/1365-2656.12013

Dingemanse, N. J., & Wright, J. (2020). Criteria for acceptable studies of animal personality and behavioural syndromes. Ethology, 126(9), Article 9. https://doi.org/10.1111/eth.13082

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ, 2, e616. https://doi.org/10.7717/peerj.616

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods in Ecology and Evolution, 5(9), Article 9. https://doi.org/10.1111/2041-210X.12225

Martin, J. S., Massen, J. J. M., Šlipogor, V., Bugnyar, T., Jaeggi, A. V., & Koski, S. E. (2019). The EGA+GNM framework: An integrative approach to modelling behavioural syndromes. Methods in Ecology and Evolution, 10(2), Article 2. https://doi.org/10.1111/2041-210X.13100

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. Journal of The Royal Society Interface, 14(134), Article 134. https://doi.org/10.1098/rsif.2017.0213

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

O'Dea, R. E., Noble, D. W. A., & Nakagawa, S. (2022). Unifying individual differences in personality, predictability and plasticity: A practical guide. Methods in Ecology and Evolution, n/a(n/a). https://doi.org/10.1111/2041-210X.13755

Schielzeth, H., & Nakagawa, S. (2022). Conditional repeatability and the variance explained by reaction norm variation in random slope models. Methods in Ecology and Evolution, 13(6), 1214–1223. https://doi.org/10.1111/2041-210X.13856

Reviewed by Aparajitha Ramesh, 07 October 2022

Report on "Repeatability of performance within and across contexts measuring behavioral flexibility" by McCune et al., 2022

The study aims to test whether behavioral flexibility in terms of reversal learning is repeatable within a general context of flexibility of behavior. In addition, they also motivate their current study with the fact that construct validity, which is whether the trait of interest is actually being tested is captured by the experimental methodology. I very much enjoyed reading the pre-registration and the current paper and commend the authors on their endeavour towards open science. Furthermore, I very much liked the explanation on how the present study is integrated into a larger study and the details including codes for analyses are also made transparent. Overall I think this paper is an excellent example of good practises in science and also addresses an interesting question.

The results stemming out of this study are also interesting for research lines on animal personality. Although my impression of the manuscript is quite positive, I have two major concerns namely (1) The focus of the study (validating methodology or testing the biological phenomenon of behavioural flexibility) (2) Repeatability estimates for temporal repeatability (low and overlaps with zero). Other than these, I have a few minor suggestions which may be useful.

(1) The focus of the current study: At several points in the introduction, emphasis is made on the methodology used for testing behavioral flexibility and testing construct validity to ensure the method is testing flexibility regardless of state, personality etc. (Eg. On line 60). However, I also see that the authors have expectations on the repeatability of behavioral flexibility as a biological phenomenon. It is therefore confusing to me how the same data is used to address both questions and whether it can delineate the two. For example, if we find that behavioral flexibility are indeed repeatable over time (significant proportion of within individual variation), does temporal repeatability give an account of consistent variation in inherent flexibility and learning in animals or does it validate the methodology used for testing? I do believe that a philosophical discussion on this is useful and methodological validation can be obtained also by relating to the behaviors in the wild. But I would appreciate if there is one clear line of thoughts here.

(2) Temporal repeatability in Table 1: Temporal repeatability is quite low. Moreover, the credible intervals overlaps with 0, which one would interpret as not significantly different from 0. So I was wondering how p.value shows that it is significant.

In line 247, it is explained that the significance of repeatability was calculated by comparing to simulated data where birds performed randomly within each reversal – Is this the same as testing when confidence intervals for R is calculated by bootstrapping and statistical significance is tested against the null hypothesis that R=0 (as done in rptR package, Stoffel et al. 2017)?

Additionally, the p.value in the main text (line 250, p=0.001) is also different from the p.value denoted in the table 1 (p=0.01).

Overall, it may be important to verify whether the repeatability value is large enough and if it is different from zero to determine whether P3a alternative should be worked out or alternative explanations are needed.

Minor comments:

* Title: Title should be changed to address the focus (see comment (1)). If this is methodological, it would be nice to explicitly have it in the title for greater comprehension

* I am not familiar with this but maybe keywords are missing?

* Line 29: Add scientific name

* Line 38: Maybe a personal preference, but perhaps the starting of the paragraph can be improved and sentence can be made shorter.

*Line 80: Scientific name should be italicised

*Line 80: 'First' is not followed by 'Second' or 'Then' – maybe remove it if the hypotheses are not discussed here but later.

*Line 117: Secondly should be preceded by 'Firstly'. Here, the Secondly describes the hypothesis mentioned in the introduction. Could the two hypotheses that are being dealt with in this paper be made clearer? Also in Line 80 (see previous comment)

Reference:

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods in Ecology and Evolution, 8(11), 1639-1644.