In this paper, the authors test whether and how measures of cognitive flexibility relate to other (repeatable) behavioral traits such as those typically interpreted as “exploration” or “boldness” using grackles. To do this, they first tested for repeatability in these other behavioral traits and found that no measures of “boldness” (approach to putatively threatening objects) were repeatable. However, they did find that some behavioral measures in the exploration assay were repeatable: birds were highly repeatable in their latency to approach and spent time near a novel environment, but not novel objects. They also found that the birds were modestly repeatable in the number of touches they made to all the different objects across the assays.

Then once the authors established that certain behaviors were repeatable, they looked for relationships between those behaviors and specifically ‘cognitive flexibility’. They found no relationship between flexibility and the exploration measures. They then also looked for relationships between persistence and flexibility and found that more flexible birds were more persistent in some ways, but not other ways.

Altogether, I actually had a bit of a hard time reading the manuscript. I really genuinely appreciate the fact that authors are clearly making a big effort to use research best practices in pre-registering their methods and questions. However, I don’t know if this is what is expected with the post-study write-ups, but this manuscript is pretty difficult to evaluate as a stand-alone document. I’m not sure if the general expectation is that any reader would first read the pre-registration and then read the post-study write-up (i.e. manuscript) but that seems to be necessary here in order to evaluate this manuscript as many of the important methodological details are missing.

So my biggest comment is that I think the authors should really put the relevant parts of the methods that were included in the pre-registration into the manuscript here so that the work can be evaluated in its entirety with just one document. The authors can describe their methods as planned and then explain when/how they deviated from these plans as necessary.

- **Title and abstract**
  - Does the title clearly reflect the content of the article? X Yes, [ ] No (please explain), [ ] I don't know
  - Does the abstract present the main findings of the study? X Yes, [ ] No (please explain), [ ] I don’t know
- **Introduction**
  - Are the research questions/hypotheses/predictions clearly presented? X Yes, [ ] No (please explain), [ ] I don’t know
  - Does the introduction build on relevant research in the field? X Yes, [ ] No (please explain), [ ] I don’t know

Just wanted to mention that I really appreciate the motivation of the study – to disentangle the jingle-jangle fallacy with these behavioral measures. This is something I think about a lot – we interepret different behaviors (e.g. latency to approach an object) as different/similar things ("exploration" versus "boldness") which makes major assumptions about what we think are the underlying causes of variation in those behaviors. I really would love to see more work that comprehensively tries to figure out what the actual axes of behavioral variation are as opposed to a priori assuming we know them!

- **Materials and methods**
    - Are the methods and analyses sufficiently detailed to allow replication by other researchers? [ ] Yes, X No (please explain), [ ] I don't know
    - Are the methods and statistical analyses appropriate and well described? [ ] Yes, X No (please explain), [ ] I don't know

1 - Altogether, I actually had a bit of a hard time reading the manuscript. I really genuinely appreciate the fact that authors are clearly making a big effort to use research best practices in pre-registering their methods and questions. However, I don't know if this is what is expected with the post-study write-ups, but this manuscript is pretty difficult to evaluate as a stand-alone document. I'm not sure if the general expectation is that any reader would first read the pre-registration and then read the post-study write-up (i.e. manuscript) but that seems to be necessary here in order to evaluate this manuscript as many of the important methodological details are missing.

So my biggest comment is that I think the authors should really put the relevant parts of the methods that were included in the pre-registration into the manuscript here so that the work can be evaluated in its entirety with just one document. The authors can describe their methods as planned and then explain when/how they deviated from these plans as necessary.

For example:

- At the start of the methods it would be super helpful to have some sort of overview – how many birds were used in total, how many times were they measured, in what behaviors? Reading the MS, I think that 19 birds were used total, and they were measured twice in the 'boldness' and 'exploration' assays but I'm struggling to figure out how/when motor diversity/persistence was measured and how many times cognitive flexibility was measured? Were these assays performed on the same days? Different days? How long were the birds in captivity?
- Line 281: For the repeatability analyses, what other fixed effects were included in the models? The authors report the (I think, adjusted) repeatability estimates so having the authors also report the marginal and conditional R-squared values from their models would be super helpful. E.g. a repeatability of 0.85 seems high, but if the fixed effects of the models are actually explaining 80% of the variance in the data, then this means that individual differences only accounts for 85% of the remaining 20% of the data.
- Line 290: How exactly did you test if these behaviors were 'correlated' with flexibility? Did you just run a simple pearsons (or spearmans) correlation? Or did you use a bivariate model to account for the repeated measures on the individuals (this would be the most appropriate thing to do).

I recognize that much/all of this information may be in the pre-reg but it really also needs to be included here.

2 – I recognize that the authors pre-registered this so their hypotheses/predictions already went through review, but I just wanted to note that the hypotheses they listed in the manuscript are actually not really (biological) hypotheses. Biological hypotheses are explanations for why something is the way it is. So for example, hypothesis 1 simply states that behavioral flexibility will be correlated with exploration but not boldness, but doesn't say why. The why bit is the hypothesis and would be the interesting bit to understand! The authors do have hints of hypotheses in some of their alternative predictions (e.g. P1 alt 4 states "[no correlation between exploration and flexibility may happen because]…these measures of exploration incorporate novelty and thus measure boldness rather than exploration." This bit here is an EXPLANATION for why there may or may not be a relationship between these measures. The second halves of P6 alt 1, 2, 3 also contain potential hypotheses. So really the things that authors have written as hypotheses here are in fact just statistical null hypotheses which are not biologically that interesting (e.g. H0: there is no relationships between X and Y; H1: there is a relationship between X and Y) but are not biological explanations for why that relationship is happening. I just mention this because I see this as a good opportunity for the authors to make their paper stronger and more impactful!

- **Results**
  - In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [ ] Yes, X No (please explain), [ ] I don't know
  - Are the results described and interpreted correctly? [ ] Yes, [ ] No (please explain), X I don't know

1 - This is something the authors can't control and I'm sure they're aware of it. But they are asking a lot of 19 data points really… For the number and complexity of analyses they are doing, it seems unlikely they have the power to detect anything but the strongest effects.

2 - I just felt like overall there was very little attention paid to the differences between the experimental and control birds? Like the authors report the overall mean level differences in their behavioral measures, but it'd be really nice to see the figures color coded as well to indicate which birds were in which group. Along these lines I also couldn't tell whether the other behavioral measures (boldness, exploration) were conducted before or after the manipulation? This seems really important for interpretation?

- **Discussion**
  - Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? X Yes, [ ] No (please explain), [ ] I don't know
  - Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [ ] Yes, [ ] No (please explain),X] I don't know