Review of HB 2016 475

Decision: REJECT

Comments to the Authors:

The authors of "Individual Fecal Corticosterone Variation Correlates with Urban-Rural Personality Trait Differences in Northern Cardinals (*Cardinalis cardinalis*)" explored the correlation of fecal CORT metabolite concentrations and behavior and compared these two measurements and their relationship to each other in two populations (two "urban" and two "rural" populations). The fecal CORT measurements were collected at three time points per bird: at capture, on day two of captivity and on day eight of captivity, which was intended to represent 'baseline' and 'short term' and 'long term' stress responses. The five behavior assays performed were aimed at estimating variation in spatial and object neophobia, risk-taking and aggression. The authors report that more proactive personalities had lower baseline and short-term stress CORT levels and that urban birds had overall lower CORT levels than rural birds. These results are interpreted to support the idea that urban environments select simultaneously (correlated selection) on proactivity and lower CORT. I applaud the authors on the large amount of effort required to do work like this with wild songbirds and I think this question of urbanization effects on endocrine and behavioral traits is an interesting one. That said, the current ms suffers from some of the same limitations and pitfalls shared by some of the previous studies addressing this question. Principal among them is replication. I applaud the authors for examining MORE than one population of urban and of rural birds, but two populations is still insufficient. The question these (and other authors) are addressing is a population biology question, and it requires thinking like a population biologist. Namely, that the unit of replication is populations, not individuals. As a separate but related point, it is puzzling that the authors did not depict (figures) and describe (results) the trait variation for each of their two replicate populations. The reader is left to assume that the two rural populations, for example, were similar to each other. Below I outline my specific comments and concerns, and I hope the authors find them helpful.

——specific comments——

Line 105: the use of 'predict' here seems out of place since it suggests that you screen for behavior in order to estimate hormone concentrations, which would be a very low power and unreliable method for estimating endocrine variation (also, why not just sample GC variation if that is the question?). Further, this is a correlational study, so behavior isn't a 'predictor' of hormone concentrations per se, but rather a correlate (e.g. the hormone-behavior relationship is likely bi-directional).

Line 253-264: These sentences highlight what seems like the major problem with this line of research, namely the lack of replication. The authors emphasize that the existing research on this topic has shown inconsistent results and then they cite two publications that do not find significant baseline CORT differences — both cited studies that had effective sample sizes of 1

(one rural, one urban 'population'; the current ms has a sample size of 2). Several of these authors have pointed out that what needs to be done to advance this question is replication (e.g. within a species, sample multiple rural and multiple urban populations that do not exhibit gene flow). The present study appears to largely repeat this common pitfall, namely pseudo-replication — i.e. the unit of replication for a question like this is the number of populations, not the number of subjects in the populations as was done here.

Line 287: This transition doesn't make sense conceptually. The authors argue here that given that there is no consensus on the urban-rural gradient for GCs, it would be more useful to also include measurements of behavior. Why? If there is a lack of consensus on the GC gradient, that may very well be driven by the problem outlined in the comment above. Adding more traits to the analysis does not necessarily help solve this problem. The only way I can imagine this helping to resolve these questions is if populations differ in the distributions of personalities (or other GC correlated traits) and these previous studies had sampled their populations in a biased way. If this is the idea, it should be made explicit.

Behavior tests 2 & 3: the difference between these two tests needs clarification.

Line 718: "to be repeatable within-individuals and consistency varied between individuals." The phrasing is redundant. Repeatability is the quotient of variances: (among individual) / (among + within); thus the authors can simply state that "each of the five behavioral measures were significantly repeatable". Furthermore, these repeatability estimates (and 95% CI) should be reported here for the reader to evaluate.

Lines 824-844: the authors performed separate analyses of 'baseline' and 'stress-induced' GCs and then an analysis on the difference between these two. These "difference in cort" analyses are highly problematic because an effect between the two populations could be driven by ONLY a difference in baseline OR stress-induced measurements. In other words, performing statistics on statistics here is not warranted because it interferes with the inference. If the authors want to pursue something along these lines, they could perform an analysis on stress GCs using 'baseline' as a covariate. Unfortunately, because the authors did not collect multiple repeated samples (e.g. two baseline samples, two stress induced samples, etc) from the same individuals, they cannot estimate the among-individual covariance in these traits, which would be the statistical approach required to get at the question they are attempting to answer with the 'difference in cort' analyses.

PCAs: I was expecting this section to examine the relationships for performance in the five different behavioral assays, rather than reduce the parameters estimated in each separate assay. The former would be in line with the concept of behavioral syndromes/ personalities (e.g. more aggressive birds are also more neophilic). Presently it appears that the ms treats these five different behavioral assays as separate traits, in which case there is no demonstration of personality per se. The authors indicate on line 718 that each of these five behaviors were repeatable, but 'personality' is conventionally defined

as repeatable and co-varying traits.

Tables: substitute ‘extraneous’ in place of ‘confounding’ (hopefully those four variables are not actually confounds!).

Figures 2-4 and the Results sections for these data: Each of the two populations of rural and of urban birds should be depicted in these graphs rather than pooled together. In my view, this replication aspect is the most important contribution of this type of work, and it is currently hidden from view. Do the two rural populations look alike for behavior, hormones and the relationships between them and likewise for the two urban populations?

Line 797: this issue has to be validated in each species. The cited studies represent work in other species.

Line 810: again, this is not known for cardinals and the results of work in doves and cranes cannot be assumed to be representative of the study species here.

Line 861: why were fecal subsamples “approximately” 0.2 grams? Why not exactly 2 grams?

Line 917, sentence beginning with “Serial” needs syntax correction.

Line 922: indicate how parallelism was evaluated statistically (test statistic, df).

Line 925: do the authors mean “samples were randomized across and within plates”? This paragraph has a number of syntax and grammar errors that need attention.

Line 933: How CVs were calculated is not clear. It sounds as though the CVs for the duplicates of the standards (the 5 standards used in the Enzo standard curve) were used. If this is the case, it would very likely greatly underestimate error considerably. Conventionally, a stripped pooled sample is spiked with a known concentration of commercial hormone and then processed two or three times (in duplicate) on each plate, and then inter and intra-assay CVs are calculated from these estimates. It also doesn’t make any sense to cite other avian studies for inter and intra-assay CVs, since this is principally a measure of pipetting error.

Line 964: the authors state here that log transformations were applied “when necessary or appropriate” for normality of each variable. It is important for their parametric analyses that the residual error from their models is normally distributed, not their input variables.

Line 1317: the authors describe “between-individual behavioral variation” but this was not measured. Each behavioral assay was only conducted once, so it is not possible to parse variation at the individual level. The variation is merely sample variation.

Additional Comments to the Chief Editor only: I see this ms as representing an incremental step — it examines a question (relationship between personality, urbanization and GCs) that has been studied in other bird species and comes to similar major conclusions. Compared to some of these previous studies, this ms suffers from a lower quality hormone assay (fecal metabolites) for the given question, less replication and some incorrect statistical approaches—e.g. the authors analysis on the difference between their cort measures (see comments to authors). There are other problematic analysis issues, including the application of several Pearson and Spearman correlations on CORT vs CORT and then each CORT measure versus each behavioral PC, etc. The probability of type 1 errors is quite high with this kind of exploratory analysis approach. Finally, the authors refer to and treat the dataset as though it examines 'between-individual variation" but in fact they only measure each trait one time per subject, so it is not possible to draw inferences about variation at the individual level (only sample variation). Likewise, personality per se was not demonstrated in this study (a common problem in this field, but one in desperate need of correction).