I've read with interest this preregistration submitted by Logan and colleagues to PCI Ecology. First, I commend the authors for submitting a pre-registration of their research project, something that is rare in ecology and evolution (I should probably get around it myself!). Although it is beyond the remit of this review, I want to say the broader project seems especially interesting, and that the current pre-registration seems like a well thought-out part of it.

Regarding the pre-registration itself: As far as I can tell, the broad methods used to answer the questions at hand are sound and appropriate, the hypotheses (and alternatives!) are well-defined, and the broad details of the statistical analyses (which dependent and independent variables?) correct. There are however several comments regarding the sample size and statistical analyses themselves that need to be addressed before recommendation:

(Remark: I address some issues in the first test where they are encountered, but they may also be present elsewhere in the pre-registration).

1- In "data checking", please note that data normality can only be assessed on residuals relative to a model, not on actual data (well, it can, but it doesn't mean anything with respect to analysis validity). Please also note that normality diagnostics plots can be extremely misleading and difficult to interpret for non-Gaussian GLM(M)s. I advise to look at the DHARMa R package (https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html) by Florian Hartig, which contains several diagnostic plots and tests that don't have this problem.

2- At several points throughout the registration, you plan to use generalized linear models (not mixed) when your list of independent variables clearly indicate a random effect of ID (see P1, but not only). If there are repeated individual measures, a mixed model should be used, even if the random effect is not what is tested in the current hypothesis.

3- I have doubts about the ability of the chosen power analyses to correctly capture the true power of your tests, especially in cases involving random effects, where data points won't be independent. I suggest the authors use simulations to get a better handle of how their models will actually behave. I acknowledge that simulating various effect sizes, for different data structures, can be time-consuming and difficult. I believe the package SQuID may be useful here (Allegue et al., 2017)(https://cran.r-project.org/web/packages/squid/index.html) (disclaimer: I have not used it yet, so can't vouch for its actual usefulness here).

4- In P1 (and elsewhere), since there are only two models ("base" and "test"), there is no need to use Akaike weights. Independently of debates around their usefulness, they are meant/most useful to compare more than 2 models. With two models, you can simply use the AIC/ BIC/… and see which model is best supported. If I remember my Burnham and Anderson correctly, a deltaAIC>4 when there are two models is equivalent to a weight of >90% given to the best model anyway.

5- In P2, you use the average response as your dependent variable. Why not include all values and an individual random effect?

6- In P2 also, please be aware that (in my experience) latencies will probably need transformation (I expect log) to fit within the statistical model.

7- In P3a, you use a linear mixed model (lmer) to estimate repeatability when a generalized linear mixed model (glmer) should be used, as the dependent variable is a count. This would lead to potentially wrong estimates of repeatability, first because the individual variance will probably be wrongly estimated, and then because the residual/overdispersion variance is not the same for a Gaussian vs a Poisson model (Nakagawa & Schielzeth, 2010)

If you want to extract correct repeatabilities from a lme4 (G)LMM (and their 95% CI) on both the response and latent scales, you can use the rptR package (Stoffel, Nakagawa, & Schielzeth, 2017).

However, since you plan to fit your GLMM using MCMCglmm you actually don't need that at all, since the latent scale adjusted repeatability and its credible interval can simply be obtained by mod$VCV[,ID]/(mod$VCV[,ID]+mod$VCV[,units]+mod$VCV[,any other included random effect]). For the raw repeatability, simply fit a model with no covariates. For the repeatability on the response scale, see (P. de Villemereuil, Morrissey, Nakagawa, & Schielzeth, 2018; Pierre de Villemereuil, Schielzeth, Nakagawa, & Morrissey, 2016) and the QGglmm R package.

8- In P3b, I'd advise to use a multivariate mixed model rather than a univariate correlation between averages. See (Houslay & Wilson, 2018) and https://tomhouslay.com/tutorials/ for the rationale behind my suggestion and the potential problems with your approaches.

9- The paragraph F on sampling size is not clear (you mention 16 birds per treatment, then 8 birds per experiment). Also please be precise what "many more" entails in "we expect to be able to test many more"; if only for ethical reasons, and then to design a proper maximum sample size stopping rule (even an approximate one). It may also be useful to run your power analyses for different potential sample sizes, not only the minimal one.

REFERENCES CITED:

Allegue, H., Araya-Ajoy, Y. G., Dingemanse, N. J., Dochtermann, N. A., Garamszegi, L. Z., Nakagawa, S., … Westneat, D. F. (2017). Statistical Quantification of Individual Differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. *Methods in Ecology and Evolution*, *8*(2), 257–267. doi:10.1111/2041-210X.12659

Houslay, T. M., & Wilson, A. J. (2018). Avoiding the misuse of BLUP in behavioural ecology. *Behavioral Ecology*. doi:10.1093/beheco/arx023

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935–956. doi:10.1111/j.1469-185X.2010.00141.x

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. doi:10.1111/2041-210X.12797

Villemereuil, P. de, Morrissey, M. B., Nakagawa, S., & Schielzeth, H. (2018). Fixed-effect variance and the estimation of repeatabilities and heritabilities: issues and solutions. *Journal of Evolutionary Biology*, *31*(4), 621–632. doi:10.1111/jeb.13232

Villemereuil, Pierre de, Schielzeth, H., Nakagawa, S., & Morrissey, M. (2016). General Methods for Evolutionary Quantitative Genetic Inference from Generalized Mixed Models. *Genetics*, *204*(3), 1281–1294. doi:10.1534/genetics.115.186536