Animal Behavior and Cognition
ABC 2022, 9(2):217-221
DOI: https://doi.org/10.26451/abc.09.02.06.2022

Great Tits Chosen for Greatness Makes Them Representative: A Commentary on Farrar et al.'s "Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research"

Emil Isaksson[1]*, A. Utku Urhan[2], and Anders Brodin[2]

[1]Department of Biology, University of Ottawa
[2]Department of Biology, Ecology Building, Lund University

*Corresponding author (Email: jisak085@uottawa.ca)

Citation – Isaksson, E., Urhan, A. U., & Brodin, A. (2022). Great Tits Chosen for Greatness Makes Them Representative: A commentary on Farrar et al.'s "Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research." *Animal Behavior and Cognition, 9*(2), 217-221. https://doi.org/10.26451/abc.09.02.06.2022

Abstract – Studies of animal cognition struggle frequently with the question of how representative results from small samples are for a species. A recent article by Farrar et al. (2021), in this journal, highlights some of the major problems and suggests some solutions to these with cautionary examples drawn from the animal cognition literature. One such example comes from a study of inhibitory control in the great tit, *Parus major*, by the authors of this commentary. Although we recognize, and agree, that there are issues regarding representativeness in studies of animal cognition, we disagree with the use of our inhibitory control study as a cautionary example. Here, we explain why we think that our study is representative of Great tit inhibitory control. In fact, some of our arguments as to why our study is representative are in agreement with suggestions by Farrar et al (2021), e.g., comparing individuals with different levels of previous experiences in the cognitive paradigm under investigation. Moreover, we also add to Farrar et al.'s (2021) conclusion on how to approach studies with ambiguous representativeness by highlighting the importance of recognizing and discussing methodological differences in studies of cognitive ability. In summary, we do not argue against the valid points laid out by Farrar et al (2021), but discuss important nuances of the representativeness issue to also consider and, most importantly, add an additional point of scrutiny to account for in comparative animal cognition research.

Keywords – Representativeness, Methodology, Comparative cognition, *Parus major*, Inhibitory control, Sample size

Comparative animal cognition research relies on sample representativeness to make meaningful inferences. However, this criterion can be hard to fulfil, as shown in a recent discussion article "Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research" – Farrar, Voudouris and Clayton (*Animal Behavior and Cognition*, 2021, 8(2):273-295). The authors have identified several potential issues pertaining to representativeness and provide suggestions of steps that can be taken to improve this. They discuss a multispecies comparison of inhibitory control, an executive cognitive function thought to be required in a multitude of cognitive tasks by preventing animals from continuously acting on potentially detrimental impulses (Diamond, 2013; MacLean et al., 2014). One of the studies used by Farrar et al. to illustrate potential problems that can lead to non-representativeness, is written by the authors of this commentary, who showed that great tits (*Parus major*) possess a high level

of inhibitory control, which is unusual for a small passerine (Isaksson et al., 2018). The task we used to test inhibitory control is called the cylinder task. In this task, individuals are tested on their ability to restrain themselves in front of a transparent cylinder containing a desirable food item after first being trained on an opaque cylinder. In ours and many other studies, inhibitory control is measured as the proportion of trials, usually out of ten (Isaksson et al., 2018; Kabadayi et al., 2016; MacLean et al., 2014), in which the animals successfully circumvent the transparent barrier without touching it before getting the reward.

In this commentary, we do not argue against any of the points that Farrar et al. made, and we agree in general with what Farrar et al. suggest; we even believe that this article was necessary and timely. We do not, however, think that the conclusions drawn from our study were correct. Here we provide a commentary on some of the points Farrar et al. (2021) made on our study. Contrary to these authors, we argue that our study can be considered representative of great tit inhibitory-control ability also if one considers the criteria provided by Farrar et al. on how to increase representativeness of samples. More specifically, Farrar et al. compare our great tit inhibitory score with that of two more recent studies to claim that our score is not representative. The inhibitory score in all three studies was calculated as the number of trials out of ten that individuals succeeded in the cylinder task (Coomes et al., 2021; Isaksson et al., 2018; Troisi et al., 2021). In our study, trained great tits scored 80%, while untrained tits scored 60%. In comparison, untrained individuals in the two more recent studies scored around 40% (Coomes et al., 2021; Troisi et al., 2021). In addition, Farrar et al. also claim that we should have used the score of other, closely related species, or species with similar ecology, to guide the performance estimate in our study. In other words, Farrar et al. suggest that we should weigh in the scores of closely related species when estimating the scores of our great tits. There are three main reasons why we disagree that our study can be used as an example of how small sample sizes can yield data that are not representative based on these arguments by Farrar et al.: i) the birds that performed at 80% in our study are not directly comparable to the birds that scored about 40% in the two other studies, ii) there are differences in methodology between our study and the studies by Coomes et al. (2021) and Troisi et al. (2021) (henceforward referred as C. and T.), and, iii) pertaining to the second argument of Farrar et al., a major aspect of comparative cognition is investigating species differences; thus, making a performance average of, or guiding performance estimate based on, a higher taxonomic group uninformative.

Experienced Great Tits Scored 80%, Inexperienced Great Tits Scored 60%

In our study, there were three treatment groups with eleven individuals in each: 1) a control group without any prior transparency-experience in the lab, 2) "in general" transparency experienced individuals, which had access to a small transparent wall (17x17cm) in their home cage before being tested (but not any transparent cylinder) and, 3) transparent cylinder experienced individuals, which had been allowed to interact with a slightly different type of transparent cylinder than the one used in the test, before the experiment (Isaksson et al., 2018). Groups one and two were subsequently pooled as their performance did not differ significantly from each other, creating a "cylinder naïve" group ($n = 22$) compared to group three ("cylinder experienced," n=11). Group three, with transparent cylinder experience, averaged 80%, whereas the pooled, "cylinder-naïve" group scored about 60% (Isaksson et al., 2018). Farrar et al. (2021) argue that the 80% score in our cylinder-experienced birds cannot be considered as representative for the species because C. and T.'s cylinder task scores were around 40% (Coomes et al., 2021; Troisi et al., 2021). We agree that our sample size is rather small; however, directly comparing trained and untrained individuals is misleading.

This argument is surprising as Farrar et al. (2021) themselves mention a few steps to improve sample representativeness, one of which is to increase homogeneity and control. This can, for example, be achieved by training individuals to get their performance close to their theoretical maximum (Smith & Little, 2018 in Farrar et al., 2021). Past experiences can increase performance in current tasks, thus making them easier (Dukas, 2017), which has been shown to be true for inhibitory test tasks as well (van Horik et al., 2018). The difference between inexperienced and experienced individuals in Isaksson et al.'s (2018) study was significant, which suggests that training in this case, on average, improves performance by

approximately 20%. Although not subject to the intensive training regime as suggested by some authors (Leavens et al., 2019), nor having the training monitored or quantified prior to the experiment, the cylinder-experienced individuals should be considered just as representative of the great tits' performance in the transparent cylinder task-test as other studies with modest sample sizes, based on Farrar et al.'s own argument.

Interestingly, the same research team that performed the C. and T. experiments has recently tested great tits in the wild (Davidson et al., 2021). In this design, the birds were allowed to enter the experimental apparatus at free will and performed at 60%, almost identical to our cylinder-inexperienced birds. Davidson et al. used a different version of the detour task, not using a cylinder; more importantly however, the wild great tits were inexperienced with the transparent obstacle. As the great tit is known for its good learning abilities (Morand-Ferron et al., 2015), we think that cylinder-experienced birds should be expected to perform better than non-experienced ones; i.e., better than 60%. Moreover some "high success rate" individuals in Coomes et al.'s (2021) study approached 80% – the average for our cylinder trained individuals – making their results consistent with those of our experienced birds.

Methodological, Rather Than Sample and Site, Differences Could Underlie Result Discrepancies

Representativeness can be achieved by reducing the amount of unwanted variation, or noise, in the data via increased control of experiments (Farrar et al., 2021). A few methodological differences point to increased control, or at least reduced chances for behavioral noise, in the experimental design of Isaksson et al.'s study, compared to C. and T.'s studies.

The experimental device itself differs between studies; we used an 8.5cm long and 3.5cm outer diameter cylinder (Isaksson et al., 2018), whereas both C. and T. used a shorter cylinder, 3cm long x 3.5cm diameter (Coomes et al., 2021; Troisi et al., 2021). A longer cylinder could encourage a clearer expression of a detour, resulting in a higher estimate compared to a shorter cylinder where the birds are not required to detour more than a step (Kabadayi et al., 2018). This explanation is based on the fact that the detour itself should have been learnt during the opaque cylinder training phase. Differences in cylinder size likely account for part of the variation in results between studies.

C. and T. tested great tits in their home cages and, in addition, tested individuals in ten consecutive trials in the same day not accounting for satiation (Coomes et al., 2021; Troisi et al., 2021). This method could influence the results by 1) the human presence while placing the cylinder in the birds' home cages is likely stressful for the individuals, which can increase the likelihood of accidental failures in the task due to irrelevant behavior, and 2) following the commonly used protocol from MacLean et al. (2014) with ten consecutive trials will increase the risk of satiation, and by extension, decrease motivation to get the food reward, which also could negatively influence the results. In contrast, we allowed birds to spontaneously move over to a testing cage with an attached, specially designed, testing box that controlled the approach of the bird to the cylinder and performed only five tests/day to avoid satiation (Isaksson et al., 2018). We feel that our design and procedure make the detour score more meaningful as accidental touches are minimized compared to placing a cylinder in the birds' home cages.

Summarily, there are five methodological differences between Isaksson et al.'s study and C and T's studies of inhibitory control in great tits in the cylinder task that could influence the score differences: Our study, 1) used a separate testing cage, 2) used a testing box that placed individuals perpendicular to the cylinder at the start of the test, 3) minimized human interaction before starting the test, 4) used a longer cylinder that should require a clearer detour response, 5) aimed to minimize satiation effect with five trials/day.

Species Comparison, not Family Mean

Interspecific comparisons of cognitive abilities are needed to answer proximate and ultimate questions about cognition (MacLean et al., 2012; Shettleworth, 2009). We do agree with the general notion that incorporating the cognitive performance of closely related species can be useful in explaining evolution

of cognitive traits. However, we do not agree with Farrar et al. (2021) that we should have incorporated performance estimates from similar species to adjust the great tit performance estimate to, a general Passeriformes mean value. An important purpose of our study was to add great tits to the inclusive data set on species performance in the cylinder task compiled by MacLean et al. Another purpose was to scrutinize the conclusion of MacLean et al. (2014), that absolute brain size should be the best predictor of cognitive performance in general. More specifically, we believe that we could investigate this by testing a bird with a relatively small absolute brain size. In particular, we were interested in the great tit because it has been shown to perform well in other cognitive tests (Brodin & Utku Urhan, 2014; Cole et al., 2011; Morand-Ferron et al., 2015; Sasvári, 1979).

Whereas a mean score on a higher taxonomic level such as family can be informative in some contexts, there is also risk of under or overestimating the mean performance that stems from fundamental ecological, behavioral and life history differences among the members of a family. Such differences may lead to closely related species being specialised in different cognitive traits. The Paridae family is very diverse (Johansson et al., 2013), and the family consists of specialist, generalist, and semi-generalist foragers and such differences can cause marked differences on performance on cognitive tests (Urhan, 2017). For example, blue- (*Cyanistes caeruleus*) and marsh- (*Poecile palustris*) tits performed much lower than great tits in a social learning task (Sasvári, 1979). Consequently, grouping the scores of these species in another cognitive task, like the detour-reaching task, would confound rather than increase the accuracy of the results. Obviously, differences in the scores of closely related species are of special interest when the aim is to understand how ecological differences between species will affect differences in cognitive ability. This is of course not to say that a family mean approach is not a valid approach as related species are more likely to share cognitive traits and more likely to perform similarly in a specific task. But such an approach should be used cautiously as the possibility that related species possess diverse cognitive abilities usually has to be taken into account.

Concluding Remarks

In conclusion, we believe that Farrar et al. (2021) make many good points that should be taken into consideration in future studies, but we think that our study is representative of inhibitory control in the great tit even though the sample size is small and, consequently, the use of our study as a cautionary example is misleading on several accounts. We think that an appropriate experimental design can make results based on small sample sizes representative of the cognitive concept under investigation. Also, we want to emphasize the importance of evaluating the methodological differences when comparing studies of cognitive concepts within species. Furthermore, we believe the concept of using experienced/trained individuals to increase sample homogenisation via experience/training when comparing species performance in cognitive tasks to avoid hidden confounds (Smith & Little, 2018 in Farrar et al., 2021), deserves merit. In our study, we had three training regimes prior to the tests - thus fulfilling the criterion proposed by Farrar on how to increase the representativeness of small sample sizes. Furthermore, we used a methodological set-up that should reduce behavioral noise compared to the other studies on great tits. Thus, although the 80% score comes from a subsample of only eleven birds, this score was chosen because we believe that it is representative and that the experience of these birds is similar to that of other animals that have been tested (Kabadayi et al., 2016).

Adding to Farrar et al.'s conclusions regarding the importance of considering sampling discrepancies between studies and the caution needed when using small sample sizes in cognitive research, we bring forth additional aspects to take under consideration; in particular, the importance of recognizing and discussing whether sampling and/or methodological differences are influencing the results. In addition, in comparative cognition, the differences between closely related species are often more interesting than the similarities.

Conflict of interest: The authors declare no conflict of interest.

References

Brodin, A., & Utku Urhan, A. (2014). Interspecific observational memory in a non-caching Parus species, the great tit Parus major. *Behavioral Ecology and Sociobiology*, *68*(4), 649–656. https://doi.org/10.1007/s00265-013-1679-2

Cole, E. F., Cram, D. L., & Quinn, J. L. (2011). Individual variation in spontaneous problem-solving performance among wild great tits. *Animal Behaviour*, *81*(2), 491–498. https://doi.org/10.1016/j.anbehav.2010.11.025

Coomes, J. R., Davidson, G. L., Reichert, M. S., Kulahci, I. G., Troisi, C. A., & Quinn, J. L. (2021). Inhibitory control, exploration behaviour and manipulated ecological context are associated with foraging flexibility in the great tit. *Journal of Animal Ecology*, 1–14. https://doi.org/10.1111/1365-2656.13600

Davidson, G. L., Reichert, M. S., Coomes, J. R., Kulahci, I. G., de la Hera, I., & Quinn, J. L. (2021). Inhibitory control performance is repeatable across years and contexts in a wild bird population. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2021.07.15.452478

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dukas, R. (2017). Cognitive innovations and the evolutionary biology of expertise. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160427. https://doi.org/10.1098/rstb.2016.0427

Farrar, B. G., Voudouris, K., & Clayton, N. S. (2021). Replications, comparisons, sampling and the problem of representativeness in animal cognition research. *Animal Behavior and Cognition*, *8*(2), 273–295. https://doi.org/10.26451/abc.08.02.14.2021

Isaksson, E., Utku Urhan, A., & Brodin, A. (2018). High level of self-control ability in a small passerine bird. *Behavioral Ecology and Sociobiology*, *72*. https://doi.org/10.1007/s00265-018-2529-z

Johansson, U. S., Ekman, J. B., Bowie, R. C. K., Halvarsson, P., Ohlson, J. I., Price, T. D., & Ericson, P. G. P. (2013). A complete multilocus species phylogeny of the tits and chickadees (Aves: Paridae). *Molecular Phylogenetics and Evolution*, *69*(3), 852–860. https://doi.org/10.1016/j.ympev.2013.06.019

Kabadayi, C., Bobrowicz, K., & Osvath, M. (2018). The detour paradigm in animal cognition. *Animal Cognition*, *21*(1), 21–35. https://doi.org/10.1007/s10071-017-1152-0

Kabadayi, C., Taylor, L. A., von Bayern, A. M. P., & Osvath, M. (2016). Ravens, New Caledonian crows and jackdaws parallel great apes in motor self-regulation despite smaller brains. *Royal Society Open Science*, *3*(4), 160104. https://doi.org/10.1098/rsos.160104

Leavens, D. A., Bard, K. A., & Hopkins, W. D. (2019). The mismeasure of ape social cognition. *Animal Cognition*, *22*(4), 487–504. https://doi.org/10.1007/s10071-017-1119-1

MacLean, E. L., Hare, B. A., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., … de A. Moura, A. C. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences*, *111*(20), E2140–E2148. https://doi.org/10.1073/pnas.1323533111

MacLean, E. L., Matthews, L. J., Hare, B. A., Nunn, C. L., Anderson, R. C., Aureli, F., … Wobber, V. (2012). How does cognition evolve? Phylogenetic comparative psychology. *Animal Cognition*, *15*(2), 223–238. https://doi.org/10.1007/s10071-011-0448-8

Morand-Ferron, J., Hamblin, S., Cole, E. F., Aplin, L. M., & Quinn, J. L. (2015). Taking the operant paradigm into the field: Associative learning in wild great tits. *PLoS ONE*, *10*(8), e0133821. https://doi.org/10.1371/journal.pone.0133821

Sasvári, L. (1979). Observational learning in great, blue and marsh tits. *Animal Behaviour*, *27*, 767–771. https://doi.org/10.1016/0003-3472(79)90012-5

Shettleworth, S. J. (2009). The evolution of comparative cognition: Is the snark still a boojum? *Behavioural Processes*, *80*(3), 210–217. https://doi.org/10.1016/j.beproc.2008.09.001

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Troisi, C. A., Cooke, A. C., Davidson, G. L., de la Hera, I., Reichert, M. S., & Quinn, J. L. (2021). No evidence for cross-contextual consistency in spatial cognition or behavioral flexibility in a passerine. *Animal Behavior and Cognition*, *8*(3), 446–461. https://doi.org/10.26451/abc.08.03.08.2021

van Horik, J. O., Langley, E. J. G., Whiteside, M. A., Laker, P. R., Beardsworth, C. E., & Madden, J. R. (2018). Do detour tasks provide accurate assays of inhibitory control? *Proceedings of the Royal Society B: Biological Sciences*, *285*, 20180150. https://doi.org/10.1098/rspb.2018.0150