Revised models still do not show evidence for motivational trade-offs modulating nociception in bees

We thank Gibbons et al for their thoughtful and constructive response to our critique. We are glad to have clarified that the evidence for bee sentience, in their view, rests on the existence of a significant negative interaction term, meaning that bees become more willing to tolerate heat as the potential reward becomes more valuable than alternatives.

In their reply, Gibbons et al provide interesting new data suggesting that bees remember their experience of heat. They show that bees avoid a previously heated high-reward feeder in conditions when there was no difference in sucrose condition but approach it when there is a difference in sucrose concentration. However, while these results do show further evidence of a lasting internal representation of rewarding and noxious stimuli, they do not by themselves show an interaction effect and thus cannot demonstrate the key motivational trade-off.

A small factual point: Gibbons et al quote us as saying "The inclusion criterion removed nearly a third of bees", which they label "incorrect". We actually wrote "nearly a third of the bees *tested in the crucial 'trade-off' conditions*". This is correct. 31 bees were tested in the trade-off conditions, of which 9 were excluded from analysis (29%). The remaining 10 bees were tested in the 40% concentration condition, in which heated and unheated feeders both offered the same reward and so there was no trade-off to be made.

The remaining point at issue concerns the evidence for that negative interaction term. We used R code kindly provided to us by Gibbons et al, in which the random effects were (1+Temperature|colony/subject). We now understand that the model actually fitted in their paper used (1+Temperature|subject) and that they have now examined a third, decorrelated model (1+Temperature||subject). They say that this last gave unchanged results but removed the singularity warning. Unfortunately, we have been unable to replicate this: all the above models still give us singularity warnings (with glmer from lme4 1.1-31 in R 4.2.1).

Based, on their response, we now understand that Gibbons et al’s significant negative interaction rests on their fitting a random effect of temperature as well as a random intercept. But since each subject was tested in only two conditions, we consider two random effects to be overfitting. Since the study tested bees at only one concentration value, the design inextricably confounds individual bees’ heat aversion with how heat aversion changes with concentration (the crucial interaction effect). It is therefore risky to draw conclusions about the interaction on this overfitted model, especially when the interaction goes away with a simpler random-effects model.

The bees certainly are highly variable in their sensitivity to heat, as we can see in the Heated/40% data. Here, bees are given the choice of feeding from two feeders, both of which offer 40% sucrose, one of which is heated. 7 out of 10 bees show no evidence of heat aversion -— statistically they are as likely to feed from the heated feeder as the unheated. Only 3/10 bees are heat averse; these bees never feed from the heated feeder. This high individual variability in heat aversion is problematic for the study, given that the very premise of the motivational trade-off is that heat is noxious. Statistically, this variability forces the model to prefer adding a random factor to account for the spread, which explains why their model scores better on the Akaike Information Criterion than our simple model.

However, from a theoretical perspective, singular fits often correspond to overfitted models that may have poor power. While the recommendation is often to use maximal models, terms leading to a singularity should be removed and likelihood ratios might be inappropriate in this case, especially for categorical data (Barr et al., 2013; Bates et al., 2018; Matuschek et al., 2017). Indeed, the preprint Gibbons et al cite in their response (Bates et al., 2018) 2018) states 'The information in the data may not be sufficient to support estimations of such complex models and may result in singular covariance matrices, even when the LMM is identifiable in principle.' We think this may be what has happened here. The preprint continues, 'In this case, we need to replace the complex LMM specification by a more parsimonious one' - as we have done.

With hindsight, we would modify our recommendations for future studies to include (a) the use of a within-subjects design such that random inter-subject differences can be clearly distinguished from the fixed-effects under investigation; and (b) the choice of a noxious stimulus that is reliably aversive for most, if not all, individuals.

In summary, strong claims require strong evidence. Even if one accepted an interaction as evidence for sentience, we still feel that an interaction term dependent on overfitting the random effects, is weak evidence indeed.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious Mixed Models* (No. arXiv:1506.04967). arXiv. https://doi.org/10.48550/arXiv.1506.04967

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001