PROCEEDINGS B

Invited reply

Author for correspondence:
Francisco Garcia-Gonzalez
e-mail: paco.garcia@ebd.csic.es

THE ROYAL SOCIETY
PUBLISHING

Risk-spreading by mating multiply is plausible and requires empirical attention

Francisco Garcia-Gonzalez[1,2], Yukio Yasui[3] and Jonathan P. Evans[2]

[1]Doñana Biological Station, Spanish Research Council CSIC, c/ Americo Vespucio, s/n, Isla de la Cartuja, 41092 Sevilla, Spain
[2]Centre for Evolutionary Biology, School of Animal Biology M092, University of Western Australia, Nedlands 6009, Western Australia, Australia
[3]Laboratory of Entomology, Faculty of Agriculture, Kagawa University, 2393, Ikenobe, Miki-cho, Kita-gun, Kagawa 761-0795, Japan

We appreciate Henshaw & Holman's [1] (henceforth HH) comment regarding our original article [2]. We understand the points made by HH, but we have reservations about their applicability to our data, as explained below. Before addressing the specific details in HH's commentary, however, it is important to note that the substance of their article deals with technical aspects of our statistical analysis, not the underlying theoretical framework or the empirical design employed in our study. Indeed, we are happy to see that HH acknowledge that our work offers a valid proof-of-principle approach for studying the role that bet-hedging plays in determining the benefits of multiple mating in isolation from other factors, namely sexually selected mechanisms. Indeed, that was the main motivation of our study, rather than to specifically document the benefits of bet-hedging and sexual selection in the subject species.

To briefly recap, our study revisited the concept of bet-hedging in the context of explaining female multiple mating, identified common misunderstandings surrounding its interpretation and offered a novel experimental approach to test for its existence. Our study system, the sea urchin *Heliocidaris erythrogramma armigera*, offers remarkable levels of experimental control via *in vitro* fertilizations, thus enabling us to analyse temporal (geometric mean) fitness among females assigned *simultaneously* to a polyandrous and monandrous mating strategy. In this way, we controlled for the effects of female genotype and maternal effects on fitness outcomes, while separating 'pure' bet-hedging effects from sexually selected paternity-biasing mechanisms by manipulating the fertilization conditions (presence or absence of competition among the gametes of different males). The underlying question addressed in our paper was to determine whether the intergenerational fitness of females is higher when they follow a strategy that increases mate sampling (i.e. polyandry) compared with a non-bet hedger strategy (monandry). By simulating reproductive bouts across generations, we uncovered the potential for bet-hedging, in addition to paternity-biasing mechanisms (e.g. sexual selection), to provide increases in fitness for multiply mated females.

HH's first comment suggests that technically, Gillespie's measure may provide a more appropriate fitness measures for our data. We appreciate the suggestion here, but note that Gillespie's measure hardly deviates from our use of geometric mean fitness, which can be understood as a valid proxy of intergenerational fitness (e.g. compare, for each trait, the data in the first and third rows in the first column of HH's table 1). Second, HH acknowledge that the analysis of alternating environments of the form ABA and BAB requires complex analysis. However, in the event, HH do not carry out such analysis. Instead, they apply bootstrapping to estimate confidence intervals (CIs) on mean effect sizes using our measure of evolutionary fitness (geometric mean) and Gillespie's measure. If the main objective of HH's comment is to provide true CIs around bet-hedging effects in our data (see below), then we assume that implicit in their decision to use our or Gillespie's measure is the fact that these measures indeed yield estimates of intergenerational fitness that are good approximations of those that would be obtained with the more complex analysis. This fact, together with the very similar values for the delta statistics provided in the first column of HH's

table 1 independently of the method used, lead us to believe that the technical points made by HH on our use of geometric mean fitness can distract the reader from focusing on the question that matters: is there scope for bet-hedging benefits when females mate with multiple males?

HH's comment further suggests that the original methods in our analyses underestimate the bounds in the fitness differentials that we found empirically. In this respect, it is important to note that our proposed design yields data on a simulated multi-generational scale and that the different generations could be arranged in different orders. As such, there was no single fixed dataset generating a single effect size (difference between the geometric mean fitness of a polyandrous strategy and a monandrous strategy), but a multitude of potential outcomes (effect sizes) depending on the ordering of generations. Our original analysis therefore included the precaution of reshuffling the order of the generations to provide a distribution of effect sizes reflecting the 'universe' of potential effect sizes that could be obtained with the real data. HH are aware of this fact but suggest that bootstrapping would have been useful for estimating the uncertainty around our effect size estimates, and that null hypothesis testing could have been carried out by generating a null distribution. We respond to each of these points in turn.

To address the point regarding CIs for our effect sizes, we agree with HH about the benefits of using bootstrapping but stress that the aim of our method was to address the uncertainty arising from the fact that effect sizes could be calculated from thousands of equally plausible combinations of real data obtained within the experiments. By contrast, the focus of bootstrapping lies in addressing the uncertainty resulting from the stochastic nature of sampling. These sources of uncertainty are different, but both are important and complementary. Our paper did not provide *true* CIs on a given, fixed, outcome, but instead provided all possible outcomes. We regret not to have made this point clearer in our study and are grateful for the opportunity of clarifying this here. We agree with HH that to approximate *true* CIs on any given estimate, bootstrapping is a more appropriate method. The key point to note, however, is that in our study there was not a single particular estimate, but a full range of them as a result of multiple equally plausible outcomes. Nevertheless, HH apply bootstrapping as if there was indeed a single estimate, but they source the bootstrapped datasets not on a single 'real' dataset but on multiple datasets arising from the reshuffling of generations. In our view, bootstrapping would be more appropriate if one of the myriad of potential re-arrangements of data yielding an effect size similar to the mean effect size obtained with the re-ordering of generations was selected. Bootstrapping could then be performed on that particular dataset. One could go further and repeat this several times with other datasets to generate a mean value (or a value close enough to the mean value), and the 95% CIs could then be averaged. The same procedure could be employed for other effect sizes on top of the mean effect size.

On this point of HH's generation of effect size CIs, we are troubled by HH's statement that 'Because individual females and males appear multiple times in the original experiment, this procedure will tend to underestimate the true population variances, and hence the strength of bet-hedging effects. We nevertheless believe this pseudoreplication is unlikely to affect the results strongly' [1]. We want to clarify that this

comment about pseudoreplication applies to HH's analysis, not to our experiment or analyses, as the above statement may seem to imply to the reader. Females in our experiment represent distinct genotypes that are assayed across three generations each (this is the purpose of the design); our analysis takes this fact into account and estimates intergenerational fitness accordingly [2]. As for males, they are not used across generations or across blocks (females). The point was to mimic females that are sampling (either monandrously or polyandrously) from a series of available males in each generation; within each block and generation, a male was shared between the two mating strategies (the male mated to the monogamous strategist), but the analysis takes into account the paired structure of the data (see [2] and associated electronic supplementary material).

An additional point raised in HH's commentary was the suggestion to test observed fitness differentials against a null distribution that assumes no difference in geometric mean fitness between monandrous and polyandrous treatments. We agree with the premise of this suggestion, but stress that it is limited by the same problem described above. HH focus on just one of a myriad of potential outcomes—one that yields the mean of the distribution of effect sizes in our original study but approximates the probability to obtain this statistic by using the whole range of datasets that our design provides. This method superimposes the re-arranging of treatments for null hypothesis testing upon the re-arranging of generations, and this is bound to produce wide CIs. Here, we outline an alternative way to test single point outcomes, including the mean of the distribution of effect sizes provided in our original study. First, we suggest extracting a single dataset that provides the mean effect size (or very approximate value). We then suggest randomizing the treatments in this dataset to obtain the null hypothesis distribution from which to get the p-value for obtaining an effect at least as large as the mean effect seen empirically (i.e. using the re-ordering of generations). This could be performed several times on numerous datasets that provide the mean effect size value (or values close to this), and one could calculate the mean p, if one wanted to get a more precise p-value than that obtained with a single dataset. This protocol could be employed to test other effect sizes on top of the mean effect size.

Despite the limitations of HH's approach outlined above, HH acknowledge that large effects of bet-hedging (experiment 1) are still plausible. Indeed, the 95% CIs calculated by HH suggest caution in rejecting the hypothesis for the absence of bet-hedging effects on offspring viability in environment A, supporting our original conclusions surrounding the potential of bet-hedging to bring benefits to multiply mated females.

We set up high levels of replication for the units of analysis within each block by setting 18 independent batches of eggs per female, and measured thousands of offspring to reduce sampling variation around the measures of female fitness. This, however, compromised the number of individual female genotypes inspected, which in turn inevitably leads to an increase in the uncertainty in our conclusions at the population level. We fully concur with HH that higher levels of replication will be ideal in future empirical tests of polyandry via bet-hedging, but emphasize that the main objective of our study was to raise awareness among researchers about empirical ways to test bet-hedging ideas, rather than to carry out a definitive test of these ideas on sea urchins.

In short, we welcome HH's critique, because it generates an interesting debate about the ways to assess significance in

complex designs similar to the design that we propose in our original contribution. HH's commentary, in conjunction with our original study and the present reply may be useful for improving our capacity to test the theory in the future. After HH's remarks, we reaffirm that the main messages and conclusions in our original study remain valid. To summarize:

(1) we provide a tractable and innovative experimental approach for addressing bet-hedging theory;
(2) our empirical results suggest that the evolution of polyandry via bet-hedging should not be overlooked. There is scope for a multiple-mating strategy to provide intergenerational increases in fitness due to benefits associated with risk spreading;
(3) our results also show that sexual selection (arising from deterministic paternity-biasing mechanisms) can augment the potential benefits of multiple mating attributable to risk spreading mechanisms (which do not require reliable mate assessment); and
(4) collectively, our findings call for an increased effort in undertaking empirical tests of bet-hedging theory in ecology and evolution.

References

1. Henshaw JM, Holman L. 2015 Bet-hedging via polyandry: a comment on 'Mating portfolios: bet-hedging, sexual selection and female multiple mating'. Proc. R. Soc. B 282, 20150346. (doi:10.1098/rspb.2015.0346)

2. Garcia-Gonzalez F, Yasui Y, Evans JP. 2015 Mating portfolios: bet-hedging, sexual selection and female multiple mating. Proc. R. Soc. B 282, 20141525. (doi:10.1098/rspb.2014.1525)