Figure 1 | Distribution of *P* values from a three-way interaction. *F*-tests from a mixed linear model fit to 36 different transformed responses from Richter *et al.*¹. The *P* values from the three responses from panels a–c of Figure 1 from ref. 1 are highlighted in a darker green.

The *P* values from the three-way interaction *F*-tests (Figure 1) allow the conclusion that there are no significant three-way interactions in these data, and therefore no strong statistical evidence for the thesis, at least on the basis of this mixed linear model. This is not to say that the thesis is not true. A larger experiment would provide more power to detect three-way interactions.

As best I can tell, Richter *et al.*¹ did not include a random cage effect in their general linear model. It turns out to be a stronger source of variability than the block effect and is the actual unit of analysis for the design, so it critically influences results. Additional mixed-model results are available in **Supplementary Figures 1–3**. There are some significant strain-by-experiment and experiment-by-condition two-way interactions and very strong main effects for strain and experiment. There are not many significant results for any effect that involves condition. Heterogenized experiments, at face value, provide a broader inference space for conclusions. There also appears to be a connection with statistical modeling done in industrial quality control, in which Taguchi and other designs are used to estimate and explore effects that are robust to noise factors.

Reproducibility is a desirable property that appears to be in a separate dimension from sensitivity and specificity. The concept of reproducibility has not been well developed in the statistical literature, and so it is no wonder that debates like this have arisen. Both sets of authors are to be commended for wrestling with the relevant issues and trying to make some headway. More rigorous statistical research is needed to provide experimenters with better guidance on how to optimize reproducibility, sensitivity and specificity in their experiments.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth.2438).

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

Russell D Wolfinger

SAS Institute Inc., Cary, North Carolina, USA.
e-mail: russ.wolfinger@sas.com

1. Richter, S.H., Garner, J.P., Auer, C., Kuner, J. & Würbel, H. *Nat. Methods* **7**, 167–168 (2010).
2. Jonker, R.M., Günther, A., Engqvist, L. & Schmoll, T. *Nat. Methods* **10**, 373 (2013).

Würbel *et al.* reply: We showed that even simple forms of systematic heterogenization of study populations may increase the reproducibility of animal experiments¹. Jonker *et al.*² claim that our conclusions are unsupported by the data and are an artifact of pseudoreplication. We disagree with their assessment, which in our view is based on misconceptions and statistical errors as outlined in detail in our Supplementary Note.

Wolfinger³ argues that neither analysis is complete and proposes a formal comparison of reproducibility by calculating three-way interactions (strain by experiment by condition). This is an interesting approach but might be misleading because it tests for any difference in the pattern of the means between conditions. In contrast, our hypothesis predicts specifically that between-replicate variance is smaller in the heterogenized condition. We agree that if replicate experiments in the heterogenized condition did not differ, this could result in an interaction effect, but so would many other patterns (such as similar differences across experiments but in opposite directions). Thus, the three-way interaction is an indirect test of our hypothesis, and because it tests for many different patterns, it has low power to detect the one pattern of interest—namely, between-replicate variance. In contrast, we compared variances directly, testing for a single outcome. However, we agree with Wolfinger³ that inclusion of “cage” in the statistical model is essential. In our general linear model, “cage” is thus included as the “strain by block (experiment)” interaction effect because there was only one cage per strain and block. We therefore agree with Wolfinger³ that the low power of his indirect method may explain his failure to confirm our result. This is further supported by his final analysis using sliced *F*-tests to approximate our approach. Although Wolfinger³ concludes that these two-way tests do not provide strong support for our thesis, they confirm our main result: the lower between-replicate variance in the heterogenized condition.

In summary, we fully agree with Wolfinger³ that heterogenization aims to provide a broader inference space for conclusions and should help to identify effects that are robust with respect to noise factors. Therefore, further research to establish effective and practicable solutions is critically needed⁴.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth.2446).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Hanno Würbel¹, S Helene Richter² & Joseph P Garner^{3,4}

¹Division of Animal Welfare, Veterinary Public Health Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland. ²RG Animal Models in Psychiatry, Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany. ³Department of Comparative Medicine, Stanford University, Stanford, California, USA. ⁴Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California, USA.
e-mail: hanno.wuerbel@vetsuisse.unibe.ch

1. Richter, S.H., Garner, J.P., Auer, C., Kuner, J. & Würbel, H. *Nat. Methods* **7**, 167–168 (2010).
2. Jonker, R.M., Günther, A., Engqvist, L. & Schmoll, T. *Nat. Methods* **10**, 373 (2013).
3. Wolfinger, R.D. *Nat. Methods* **10**, 373–374 (2013).
4. Richter, S.H. *et al.* *PLoS ONE* **6**, e16461 (2011).