Does systematic variation improve the reproducibility of animal experiments?

To the Editor: Reproducibility of results is a fundamental tenet of science. In this journal, Richter *et al.*[1] tested whether systematic variation in experimental conditions (heterogenization) affects the reproducibility of results. Comparing this approach with the current standard of ensuring reproducibility through minimizing variation in experimental conditions (standardization), they concluded that heterogenization improved reproducibility[1]. However, in our view, they did not account for significant sources of dependency in their data, which resulted in an inflated type I error rate through pseudoreplication (defined as "the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent"[2]). We show that this leads to strong overconfidence in their analyses and that their hypothesis is unsupported.

Richter *et al.*[1] compared F ratios of strain-by-experiment interactions of 36 behavioral measures to test for differences in reproducibility between series of standardized and heterogenized experiments. Although these measures were treated as independent, most are strongly intercorrelated, thus causing interdependency of the F ratios. Two sources contribute to interdependence of behavioral measures and, hence, pseudoreplication. First, the measures within each of their three behavioral tests are strongly intercorrelated. For example, an animal exploring the edge of an arena cannot explore the center simultaneously. Second, the measures are correlated across tests because many animals show temporal and cross-contextual consistency in behavioral traits (known as 'animal personalities'[3,4]). However, to allow meaningful comparison of reproducibility for standardized versus heterogenized experiments, F ratios must be obtained independently[5].

We reanalyzed their data (**Supplementary Note** and **Supplementary Figs. 1–6**) by identifying supposedly independent variables using hierarchical clustering. This analysis showed that there was no detectable difference in reproducibility between standardization and heterogenization. Hence, in our view, their data do not support their hypothesis.

We caution that overconfidence resulting from pseudoreplication may lead to premature conclusions in studies designed to prove this principle[1,6,7]. Unjustifiably assuming that heterogenization yields better reproducibility may prompt a reduction in the number of replicate experiments, possibly decreasing the chance of detecting desired and/or unwanted effects[8]. Hence, further studies validating the benefits of heterogenization for reproducibility are required before it can be adopted as the new standard.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth.2439).

ACKNOWLEDGMENTS
We thank H. Würbel and H. Richter for making the data used in Richter *et al.*[1] available to us. We thank the members of the Stats Club of Bielefeld University's Evolutionary Biology, Animal Behaviour and Behavioural Biology departments; H. Würbel for discussion; and F. Schurr for helpful comments on an earlier version of this manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Rudy M Jonker[1], Anja Guenther[2], Leif Engqvist[3] & Tim Schmoll[3]

[1]Department of Animal Behaviour, Bielefeld University, Bielefeld, Germany.
[2]Department of Behavioural Biology, Bielefeld University, Bielefeld, Germany.
[3]Evolutionary Biology, Bielefeld University, Bielefeld, Germany.
e-mail: mrjonker@gmail.com

1. Richter, S.H., Garner, J.P., Auer, C., Kunert, J. & Würbel, H. *Nat. Methods* **7**, 167–168 (2010).
2. Hurlbert, S.H. *Ecol. Monogr.* **54**, 187–211 (1984).
3. Wolf, M. & Weissing, F.J. *Trends Ecol. Evol.* **27**, 452–461 (2012).
4. Lewejohann, L., Zipser, B. & Sachser, N. *Dev. Psychobiol.* **53**, 624–630 (2011).
5. Schumann, D.E.W. & Bradley, R.A. *Ann. Math. Stat.* **28**, 902–920 (1957).
6. Richter, S.H. *et al. PLoS ONE* **6**, e16461 (2011).
7. Richter, S.H., Garner, J.P. & Würbel, H. *Nat. Methods* **6**, 257–261 (2009).
8. van der Staay, F.J., Arndt, S.S. & Nordquist, R.E. *Genes Brain Behav.* **9**, 849–855 (2010).

Reanalysis of Richter *et al.* (2010) on reproducibility

To the Editor: The thesis put forth by Richter *et al.*[1] is that heterogenization improves the reproducibility of animal experiments over the more common practice of standardization. As support, they present an analysis of a somewhat complex experiment with 36 different responses and show primary results in their Figure 1. In the current issue of *Nature Methods*, Jonker *et al.*[2] pose some challenges to the thesis, one of which is the lack of standard errors in Figure 1; another is that the responses are correlated, and so it is not obvious that the difference shown in Figure 1d is significant.

Although there is some validity to claims from both sides, neither group appears to have analyzed this data set with a complete model, and so there is a lack of convincing statistical evidence regarding reproducibility. To this end, I reanalyzed the data using a mixed linear model that attempts to capture all relevant sources of variability and provides a basis for a more formal statistical test of reproducibility. The model includes a full three-way factorial and heterogeneous variances (**Supplementary Note 1**).

One way to pose the question is as follows: do the two-way differences (differences of differences) themselves differ across conditions? This can be formally tested by the three-way interaction term in the mixed linear model. If the thesis were true, we would expect to see some significant three-way interactions because the differences from the heterogenized experiments would have a more consistent pattern than would those from the standardized experiments.