The demand for replicability of behavioral result: from burden to asset

Ilan Golani[1], Yair Wexler[2] and Yoav Benjamini[2]

[1]Department of Zoology and Sagol School for Neuroscience, Tel Aviv University, Tel Aviv, Israel. [1]ilan99@post.tau.ac.il.
[2]Department of Statistics and Operatio Research and Sagol School for Neuroscience, Tel Aviv University, Tel Aviv, Israel.

Introduction

The demand for replicability of behavioral results across laboratories is viewed as a burden in behavior genetics. We demonstrate how, by using replicability as a design concept, we turn it into an asset, offering a quantitative criterion that guides the design of better ways to describe behavior. In this presentation we focus on how we use replicability for the design of better quality building blocks of exploratory behavior. Passing the high benchmark dictated by the replicability demand and at the same time obtaining measures that have sufficient discriminative power requires higher quality data. Some of the procedures we use to obtain such data include measuring behavior in reference to natural frames of reference and natural origins of axes used by the animals themselves, using individually customized cutoff points between the building blocks of behavior, and selecting those building blocks that fulfill the criteria for replicability, all the while filtering out building blocks that are not replicable across laboratories.

Sorting incursions into replicable and non-replicable types

We will demonstrate this procedure with a single example. Direct observation of mouse open field behavior suggests that arena wall is used by the mice as a natural origin of axes for the performance of incursions - forays into arena centre that start and end at the wall. Given this observation it only makes sense to use the wall as the origin of axes for the measurement of incursions. Observation further suggests that the depth of incursion into the center is a relevant parameter of this behavior. Therefore, it only makes sense to classify incursions according to the maximal distance they reach from the wall. The empirical density distribution of maximal distance from wall of incursions (see Figure 1a) suggests that the population of incursions consists of a mixture of three sub-populations. This impression is supported by fitting a Gaussian mixture model to the empirical density function (see Figure 1, from[1]).

Without a demand for replicability and a statistical yardstick for measuring it our study would end up by adding three incursion types to the mouse's repertoire ("ethogram"). A test for replicability endows us with the absolutely necessary function of examining the validity of these three novel building blocks beyond the reality of our own lab and the function of revealing differences across strains, treatments and preparations. With regard to incursions, testing for replicability reveals that only two of the incursion types are replicable across laboratories in the forced open field test (see Figure 2).

As shown, a comparison of the numbers of Incursions performed during a session in, for example, C57BL/6J and DBA/2J mice in 3 labs, shows that the number is higher in the first strain, but this difference is not statistically significant (see Figure 2, top left panel). Scoring of the three incursion types that are isolated by classifying incursions according to their maximal distance from wall (see Figure 1) is plotted in Figure 2 (top right) and two bottom panels. The numbers of near-wall incursions are evidently not replicable across laboratories: although there is a large strain difference in TAU, in the other two laboratories the two strains have a similar number of near-wall incursions. In contrast, there is a replicable strain difference in the number of intermediate and arena-crossing incursions, with C57BL/6J mice making significantly more incursions of both these types than DBA/2J mice in all three laboratories. It now becomes evident that the failure to achieve significant results in the overall number of incursions (See Figure 2, top left) is due to inter-laboratory variation in the numbers of a single incursion type: near-wall incursions. The boxplot summaries disclose how stratified scoring plus a multi-laboratory experiment transform lesser quality measures into higher-quality measures: we start with non-

Proceedings of Measuring Behavior 2014, (Wageningen, The Netherlands, August 27-29, 2014).
Editors: A.J. Spink, L.W.S. Loijens, M. Woloszynowska-Fraser & L.P.J.J. Noldus. www.measuringbehavior.org

significant differences in the un-stratified measure (See Figure 2, all incursions; partly overlapping spread of the boxplots of the two strains in all three laboratories), and continue with filtering out the segment type that shows a strong interaction (near-wall incursions), and we end up by keeping two new replicable measures, that of Number of Incursions in intermediate and that of Number of Incursions in arena-crossing incursions (that show a similar, consistent difference between the two strains' boxplots in all three laboratories).

Summary

the current outcry with regard to the lack of replicability of behavioral phenotyping results highlights only one aspect of the crisis, having to do with the poor predictive value of scientific results that reflect inappropriate handling of data[2][3]. The other aspect, also contributing to poor predictive value of results, is largely ignored. It has to do with the poor quality of the measures that are used to establish phenotypic differences. By judiciously selecting candidate measures (e.g., aspects of behavior that are suspected to be performed in reference to candidate reference values selected by the organism itself, as, for example. incursions that seem to be performed by mice in reference to arena wall), and by judicious preparation of the data for analysis (e.g., smoothing[4] and segmentation[5] based on intrinsic statistical and geometrical features of the data) we increase the likelihood of obtaining predictive results. It is, however, a proper test for replicability that validates or refutes our initial selections, improving the universal status of measures and building blocks that receive a high replicability score[6].

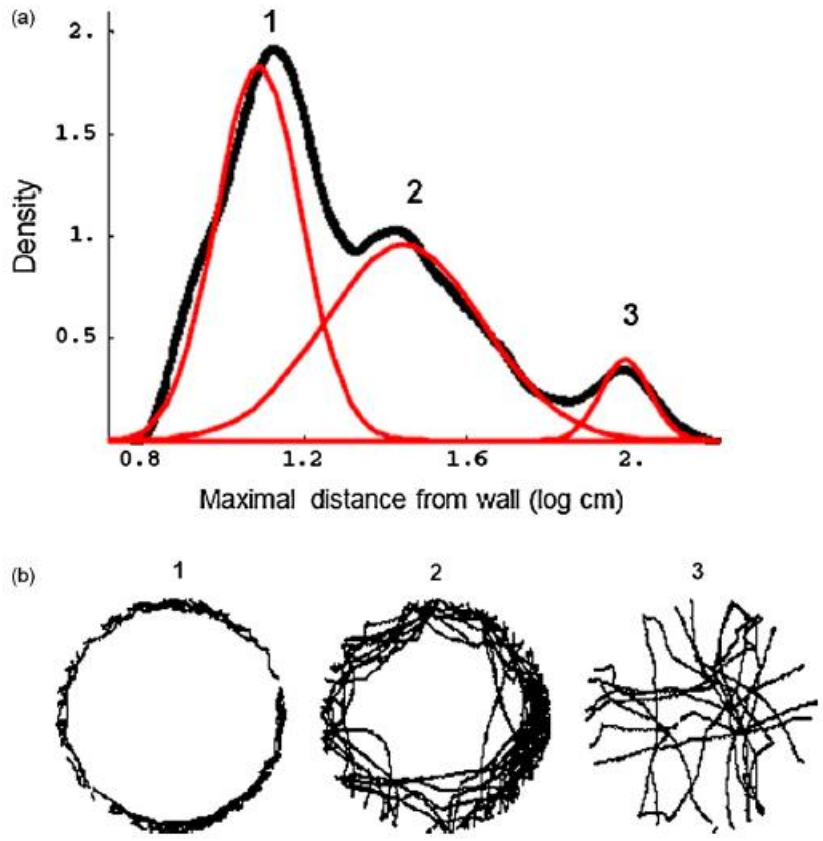

Figure 1. Black: a density graph of the distribution of the maximal distances from wall of centre segments (log transformed) in a single C57BL/6J session. Red: three Gaussians fitted to the distribution by the EM algorithm. The intersection points between the Gaussians serve as cutoff values for dividing all incursions performed in this session into three types. (b) Path plots of the incursions belonging to each type (from[1]).

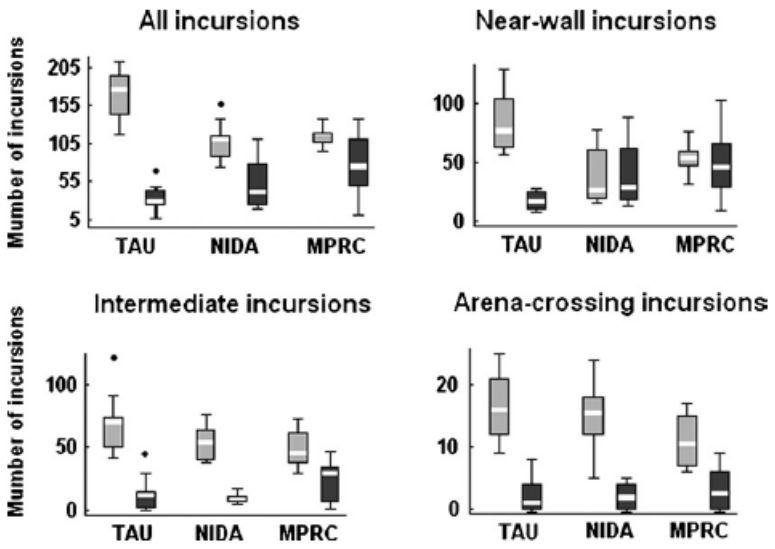

Figure 2. Comparison of the degree of replicability of the number of incursions performed during a session in DBA/2J (in gray) and C57BL/6J mice (in black) before and after stratified measurement. Top left: measurement of all incursions pooled together (before stratification). The differences between the strains are not consistent (nonreplicable) across laboratories. Top right and two bottom panels: After stratified measurement, the difference between the strains across laboratories becomes more consistent in intermediate and in arena-crossing incursions (this is represented visually in the consistent vertical distances between boxplots positions in the graphs). It is the non-replicable difference between near-wall incursions in the 2 strains that masked the highly replicable differences in the other 2 incursion types (from [1]).

Acknowledgements

This study was funded by a European Research Council grant PSARPS to YB.

References

[1] D. Lipkind, A. Sakov, N. Kafkafi, G. I. Elmer, Y. Benjamini, and I. Golani, "New replicable anxiety-related measures of wall vs center behavior of mice in the open field.," *Journal of applied physiology (Bethesda, Md. : 1985)*, vol. 97, no. 1, pp. 347–59, Jul. 2004.

[2] J. P. A. Ioannidis, "Why most published research findings are false.," *PLoS medicine*, vol. 2, no. 8, p. e124, Aug. 2005.

[3] G. Cumming, "Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better," *Perspectives on Psychological Science*, vol. 3, no. 4, pp. 286–300, Jul. 2008.

[4] I. Hen, A. Sakov, N. Kafkafi, I. Golani, and Y. Benjamini, "The dynamics of spatial behavior: how can robust smoothing techniques help?," *Journal of Neuroscience Methods*, vol. 133, no. 1–2, pp. 161–172, Feb. 2004.

[5] D. Drai, Y. Benjamini, and I. Golani, "Statistical discrimination of natural modes of motion in rat exploratory behavior," *Journal of Neuroscience Methods*, vol. 96, no. 2, pp. 119–131, Mar. 2000.

[6] N. Kafkafi, Y. Benjamini, A. Sakov, G. I. Elmer, and I. Golani, "Genotype–environment interactions in mouse behavior: A way out of the problem," *Proceedings of the National Academy of Sciences of the United States of America* , vol. 102 , no. 12 , pp. 4619–4624, Mar. 2005.