Environmental standardization: cure or cause of poor reproducibility in animal experiments?

S Helene Richter[1,2], Joseph P Garner[3] & Hanno Würbel[1]

It is widely believed that environmental standardization is the best way to guarantee reproducible results in animal experiments. However, mounting evidence indicates that even subtle differences in laboratory or test conditions can lead to conflicting test outcomes. Because experimental treatments may interact with environmental conditions, experiments conducted under highly standardized conditions may reveal local 'truths' with little external validity. We review this hypothesis here and present a proof of principle based on data from a multilaboratory study on behavioral differences between inbred mouse strains. Our findings suggest that environmental standardization is a cause of, rather than a cure for, poor reproducibility of experimental outcomes. Environmental standardization can contribute to spurious and conflicting findings in the literature and unnecessary animal use. This conclusion calls for research into practicable and effective ways of systematic environmental heterogenization to attenuate these scientific, economic and ethical costs.

Reproducibility is crucial to all laboratory research, especially in animal research where the lives of the animals are highly valuable. For example, the US animal care and use regulations require scientists not to "unnecessarily duplicate previous experiments"[1,2]. This explicitly assumes that results of animal experiments are reproducible in different laboratories and that duplication therefore represents unnecessary animal use. To guarantee reproducibility of experimental outcomes, laboratory animal science textbooks advise experimenters to standardize the conditions of their experiments. Standardization here refers to "the defining of the properties of any given animal (or animal population) and its environment" and is recommended to "increase the reproducibility of group mean results from one experiment to another",

thereby "improv[ing] comparability of results within and between laboratories"[3]. Although 'the defining of the properties' does not necessarily implicate identical environmental conditions for all animals within an experiment, environmental standardization is generally equated with such environmental homogenization. Thus, standardization renders animals within experiments more homogenous.

Standardization, test sensitivity and reproducibility

By reducing variation in the data, environmental standardization increases test sensitivity[3]. Because higher test sensitivity allows a reduction of sample size, standardization is promoted for ethical reasons also in view of reducing animal use[4,5]. That environmental standardization also increases the reproducibility of results has never been formally tested and can be repudiated on two counts. First, test sensitivity alone does not guarantee reproducibility of a result, not least because increasing test sensitivity increases the risk of false positive results[6]. Second, reproducibility is not determined by variation within experiments (as measured by sensitivity) but by variation between experiments.

Environmental standardization between experiments and laboratories aims to avoid such between-experiment variation[3,7,8]. But many environmental factors (for example, staff, room architecture and noise) cannot be equalized between laboratories so that different laboratories inevitably standardize to different local environments[7–9]. To be reproducible across laboratories, experimental results would therefore have to be applicable to at least the range of environmental conditions covered by such inherent laboratory differences. The "applicability of a result to other conditions, populations or species" is termed 'external validity'[10,11]. It refers to the robustness of a causal relation outside the narrow circumstances in which it was established[12] and thus defines the extent to which a result can be generalized. The same authors that

recommend standardization to increase reproducibility of results state that experimental results only hold for the conditions under which the experiment has been carried out, and that one can therefore not be sure whether the results can be generalized[3,13]. Nonetheless, rigorous standardization is recommended because any limitation of generalization of results owing to standardization is considered to be negligible[3,8].

Several studies on behavioral differences between inbred or mutant mouse strains cast serious doubt on this view. Thus, despite extraordinary efforts to equalize husbandry and test conditions across three laboratories, one study[7] found that some strain differences in common behavioral tests were poorly reproducible. Similar results were found in several other multilaboratory studies[14–19] and in multifactorial analyses of large datasets within laboratories[20,21] and between laboratories[22] as well as in many single-laboratory studies[23–27]. These findings indicate that even very subtle differences between laboratory, housing or test conditions can lead to conflicting test outcomes and raise the question as to whether standardization is the appropriate approach to resolve this problem.

Is standardization really the answer?

Conflicting findings between replicate studies have led to an extensive debate about standardization from which two opposing views have emerged, namely (i) that more rigorous environmental standardization will resolve the problem of poor reproducibility[8,28–30] and (ii) that standardization itself is the cause of this problem[9,31] (see also H. Würbel and J.P. Garner, Refinement of rodent research through environmental enrichment and systematic randomization; http://www.nc3rs.org.uk/news.asp?id=395).

The latter is supported by the study of phenotypic plasticity. Most biological traits exhibit environment-dependent plasticity[32], resulting in different phenotypic expressions (or states) depending on an animal's environmental background (its life history). The interaction between an experimental treatment and the animals' environmental background may therefore result in treatment effects that are idiosyncratic to that environmental background[7,24,25]. This is illustrated by the analysis of a large data archive on differences in thermal nociception between 40 strains of mice, which showed that only 27% of the variation in the data was due to strain (that is, genotype), whereas 42% was due to environment and 18% to interactions between strain and environment[20]. Therefore, the external validity of experimental outcomes inevitably decreases with increasing environmental standardization, that is, when the range of phenotypes represented in the study population becomes narrower.

Because many environmental factors resist standardization between laboratories[8], animals within laboratories will be more homogenous than animals between laboratories when environmental conditions are standardized. We therefore predict that increasingly rigorous environmental standardization within laboratories will produce results that are increasingly distinct between laboratories and hence less reproducible. This has been referred to as the 'standardization fallacy'[31]. Environmental standardization could therefore be a major cause of spurious results and conflicting findings in the scientific literature. Instead of environmental standardization, we propose that systematic environmental variation would improve the reproducibility of results, as long as the animals are 'matched' such that for each treatment animal, a control animal is selected from the same microenvironment (such as 'matched pairs' designs typical in human research). Thus, systematic environmental variation renders the animals within experiments more heterogeneous, which should increase the external validity of the results, without confounding that variation with variation owing to treatment. This prediction is particularly powerful because it contradicts current dogma in laboratory animal science, but it is difficult to test because the same experiment cannot be performed twice with the same animals.

Testing standardization against heterogenization

To solve this problem, we adopted a subsampling approach, essentially simulating two different experimental designs using data from the same real animals. We used previously published data[16] from 432 female mice of two inbred strains (C57BL/6J, DBA/2) and one hybrid strain (B6D2F1) that had been tested for behavioral strain differences in a multilaboratory study involving three different laboratories. Each laboratory had ordered three independent batches of mice that were housed in either enriched or unenriched cages (balanced across strain and batch). The resulting 18 cohorts ($n = 8$ mice per strain and per cohort) were each standardized for a unique combination of laboratory, batch and housing condition (**Fig. 1**), reflecting usual differences between replicate studies. They were thus treated as 18 replicates of a standardized experiment. These were compared with 18 replicates of a heterogenized experiment, each composed of 8 matched triplets of mice of the three strains that were pseudo-randomly selected from 8 different standardized replicates to mimic systematic environmental heterogenization (**Fig. 1**). Thus, we selected mice to either minimize ('standardized' replicates) or maximize ('heterogenized' replicates) environmental variation within replicates. To ensure that any effect was due to how, rather than which, mice were sampled, we assigned each mouse to one standardized and one heterogenized replicate, and we performed identical statistical analyses on standardized and heterogenized replicates. For the analysis, we selected five typical measurements from each of four behavioral tests that are commonly used, for example, in drug screening or behavioral phenotyping of mouse mutants (elevated O-maze test, open-field test, novel-object test and Morris water maze; for details, see ref.16).

To examine the reproducibility of the results, we first determined how variable the results were across the 18 replicate experiments. We separately analyzed each of the two experimental designs (standardization versus heterogenization) using the general linear model

$$y = \text{strain} + \text{replicate} + \text{strain} \times \text{replicate}$$

to determine the F ratios of the 'strain × replicate' interaction term for each of the 20 behavioral measures. The F ratio of the 'strain × replicate' interaction is the test statistic for the null hypothesis that there is no interaction effect. Under the null hypothesis that the effect of strain is consistent across replicates, the F ratio should equal 1. Thus, we analyzed these F ratios using the general linear model

$$y = \text{experimental design} + \text{behavioral measure}$$

(without an interaction term) to compare between-replicate variance between standardization and heterogenization across all 20 behavioral measures. To meet the assumptions of parametric analysis, we graphed and examined residuals of all behavioral measures for normal distribution, homoscedasticity and outliers, and transformed the data using angular, square-root or logarithmic transformations as necessary. The variance between replicate experiments was significantly greater in standardized replicates compared to heterogenized

replicates ($P < 0.001$), indicating that standardization resulted in poorer reproducibility of the results (**Fig. 2**).

To examine further whether this finding was associated with a higher incidence of spurious results in standardized replicates, we assessed the rate of 'false positive' (type I error rate α) results. Thus, we established the 'true' strain differences by pooling all 432 mice ($n = 144$ mice per strain), regardless of laboratory, batch and housing condition. The pooled data reflected the 'true' strain differences in the sense that they provide the best estimate of what on average different laboratories would have found. Considering that results of animal experiments should be reproducible across different laboratories, this was the best way of estimating the 'true' strain differences for the purpose of testing our hypothesis. Thus, we used the pooled data to calculate the overall strain differences for all 20 behavioral measures. Strain differences in the pooled data and in each standardized and heterogenized replicate were calculated using one-way ANOVA with strain as fixed factor and Tukey's honestly significant difference post-hoc test. Of the 60 strain comparisons, 47 yielded significant differences ($P < 0.05$), whereas the remaining 13 comparisons were not significant. We considered a result to be a 'false positive' when two strains differed significantly ($P < 0.05$) in a replicate where there was no significant difference in the pooled data. A 'false positive' result therefore reflects an idiosyncratic outcome of a single replicate experiment. Indeed, we found that standardized replicates produced 9.4% 'false positive' results (22/234 results), compared to only 1.3% (3/234 results) among heterogenized replicates (**Fig. 3a**), indicating that standardization increases the rate of idiosyncratic results. Moreover, in contrast to heterogenized replicates, the rate of 'false positive' results in standardized replicates was significantly higher than expected by chance alone ($P < 0.001$) (**Fig. 3a**), indicating that standardization may introduce a systematic source of false positive results. We thus examined whether this was merely a statistical artifact or a biologically meaningful effect.

The likelihood of obtaining a 'false positive' result depends on the statistical power of the analysis[6]. Because we expected environmental heterogenization to increase variation in the data within replicates, we also expected statistical power to be lower in heterogenized replicates. We therefore needed to make sure that the above finding was not an artifact caused by the lower statistical power in heterogenized replicates. Thus, we determined the rate of 'false negative' results (type II error rate, β), which is inversely related to statistical power ($1 - β$). We considered a result to be a 'false negative' when two strains failed to differ significantly in a replicate when there was a significant difference in the pooled data ($P < 0.05$). Owing to the small sample sizes of individual replicates ($n = 8$ mice per strain) compared to the pooled data ($n = 144$ mice per strain), the rate of 'false negative' results was relatively high. As expected, we found 58.2% 'false negative' results (492/846 results) among standardized replicates, compared to 69.6% (589/846 results) among the heterogenized replicates (**Fig. 3b**), indicating that statistical power was higher in standardized replicates. This could mean that the lower rate of 'false positive' results among heterogenized replicates indeed reflects the reduced probability of detecting significant differences.
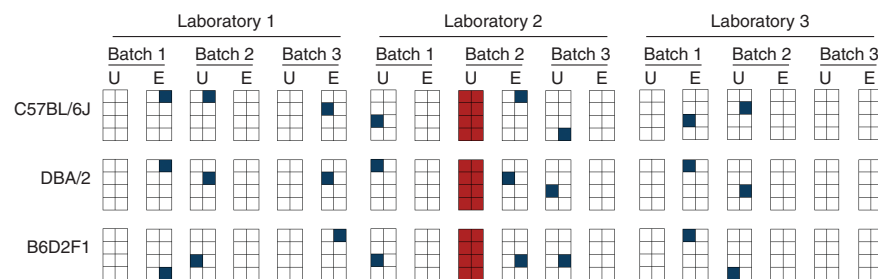

Figure 1 | Study design. In this study, 432 female mice (represented by squares) of three strains (C57BL/6J, DBA/2, B6D2F1), distributed across three laboratories, three batches per laboratory and two housing conditions (U, unenriched cages and E, enriched cages), were allocated to 18 standardized and 18 heterogenized replicate cohorts. Examples of one standardized (red squares) and one heterogenized replicate (blue squares) are displayed. Note that the heterogenized cohorts were selected such that each mouse was matched with two mice of the other two strains from the same environment.

Therefore, we reanalyzed all heterogenized replicates using a relaxed significance level of $P < 0.1$ (instead of $P < 0.05$) to equalize statistical power between standardized and heterogenized replicates. The relaxed significance level reduced the rate of 'false negative' results to that found in standardized replicates, confirming equal statistical power (**Fig. 3b**). Despite equalizing statistical power, however, the rate of 'false positive' strain differences remained significantly lower in heterogenized replicates and was not higher than expected by chance alone (**Fig. 3a**), confirming that these findings were not artifacts produced by lower test sensitivity. This demonstrates that environmental standardization introduces a systematic source of idiosyncratic results above that expected by chance alone and thus inherently fails to guarantee reproducible results.

The importance of environmental heterogenization

Our findings are based on three specific strains of female mice and 20 specific measures of behavior. House mice (from which laboratory mice are derived) are known for their remarkable
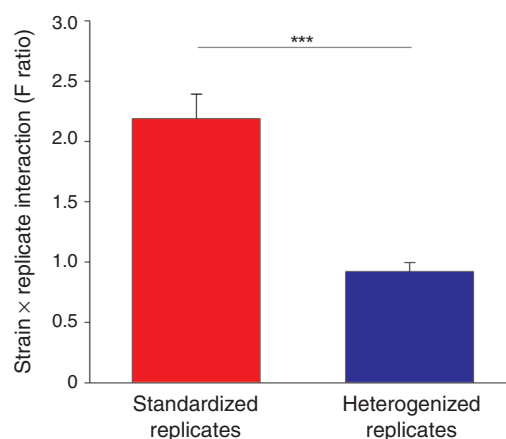

Figure 2 | Variance between replicate experiments. Variance is displayed in terms of the mean F ratio ($\pm$s.e.m.) of the strain × replicate interaction terms for all 20 behavioral measures, and was calculated separately for the two experimental designs (standardization versus heterogenization). Standardization resulted in significantly higher between-replicate variance compared with heterogenization ($F_{1,18} = 40.331$, ***$P < 0.001$). Each replicate was based on $n = 8$ mice per strain.
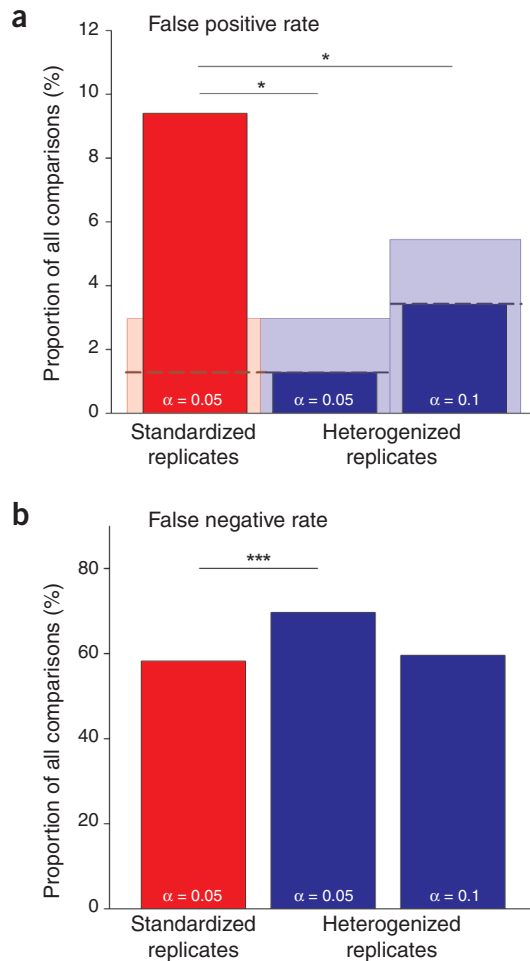
Figure 3 | False positive and false negative rate. (**a**) False positive rate (%) with respect to 'true' strain differences in the pooled data was significantly lower in heterogenized replicates (Wilcoxon signed-rank test, one-tailed, $Z = -2.24$, $*P = 0.016$), even with a critical α of $\alpha = 0.01$ for heterogenized replicates to equalize statistical power ($Z = -1.80$, $*P = 0.033$). Moreover, false positive rate in standardized replicates was significantly higher than expected by chance (binomial test, c.i. > 95%, $P < 0.001$). Dashed lines denote the expected false positive rate derived from the binomial distribution, whereby the probability of false positives was corrected for multiple testing (three strain comparisons per test measure) using the formula $1 - (1 - \alpha)^{1/3}$ to maintain family-wise error rates of 0.05 or 0.1, respectively. The light, wide bars mark the upper bounds of the confidence intervals above which the false positive rate was significantly higher than the expected rate. (**b**) False negative rate (%) was significantly lower in standardized replicates (Wilcoxon signed-rank test, two-tailed, $Z = -3.56$, $***P < 0.001$), indicating higher statistical power (power = $1 - \beta$). Statistical power was equalized by using a relaxed critical α of $\alpha = 0.01$ for heterogenized replicates ($Z = -0.45$, $P = 0.333$, not significant).

strain differences. Systematic environmental heterogenization might therefore greatly reduce the incidence of conflicting findings in the literature.

Notably, exactly the same proposal has been made with respect to genetic variation; the Banbury Conference on genetic background in mice recommended introducing systematic genetic variation into the analysis of mutant mice to increase the reproducibility of phenotypic effects[35]. Similarly, van der Staay suggested[13] "an alternative approach to increase reproducibility and generalizability, but not necessarily intragroup variation[36], is to use samples from a heterogeneous stock or from a mosaic population." Based on theoretical considerations as well as our present findings, we propose to extend the same logic to environmental variation and replace environmental standardization by systematic and controlled environmental heterogenization.

Future directions
Our findings have important implications for behavioral screening studies such as mutant mouse screens, drug screening and toxicology tests. However, as a proof of principle, their importance may reach far beyond behavioral animal research. Poor reproducibility and a lack of external validity owing to site-, study- and sample-specific idiosyncrasies occur throughout laboratory research from mass spectrometry proteomic profiling[37] to the social and behavioral sciences[38,39]. By increasing the risk for spurious results, standardization may create scientific uncertainty in many areas of laboratory research. This generates a need for replicate studies, which causes unnecessary economic costs and, with respect to animal research, undermines the ethical goal of reducing animal use by increasing test sensitivity. Systematic heterogenization of study populations or samples could attenuate these scientific, economic and ethical costs.

Further research is needed, however, to transfer the principle presented here into practice. It would be neither practicable nor efficient to design all animal experiments as multilaboratory studies. What is needed are methods for within-laboratory heterogenization resulting in populations that better represent the range of environmental variation between laboratories. The age of the animals and various aspects of their housing conditions (for example, cage size, type of enrichment and group size) are promising variables in this respect. Systematic variation of these variables is practicable and compatible with most studies, and all variables mentioned above have demonstrated effects on a wide range of potential outcome measures[20,24,40–47]. Combined with targeted experimental designs

phenotypic plasticity[33], which allowed them to successfully follow humans around the globe[34] and also predestined them to become laboratory animals. Moreover, behavioral traits are naturally more plastic than most other phenotypic traits. Whether our findings generalize to other populations and species, and to other outcome measures, are empirical questions. As a proof of principle, however, these findings demonstrate that environmental standardization may compromise the reproducibility of behavioral strain differences by systematically increasing the incidence of results that are idiosyncratic to study-specific environmental conditions. This is likely due to non-additive interactions between genotype and environment resulting in local 'truths' with little external validity[9,31]. Our results may even underestimate the true extent of the problem as test sensitivity of individual replicates was rather low because of small sample sizes.

Furthermore, our findings demonstrate that the reproducibility of results may be improved, and spurious results avoided, by introducing adequate environmental heterogenization into the experimental design. A recent literature survey on the stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades underscores the importance of our finding. Although strain differences in some measures (for example, alcohol preference and locomotor activity) were fairly robust, others (such as social behavior and measures of anxiety) varied inconsistently between laboratories and across decades[22]. The latter represent spurious results that could possibly have been avoided by systematic environmental heterogenization, without compromising the detection of robust

(for example, matched-pairs, split-plot or randomized block designs) and adequate analytical techniques (for example, matching, blocking, stratification and partialling), environmental heterogenization may be implemented in systematic and controlled ways, and without reducing test sensitivity and statistical power[6,36]. Environmental heterogenization might thus contribute to the refinement of animal experimentation in the best of meanings of the three 'R' (reduction, refinement and replacement) concept[48].

1. US Department of Agriculture, Animal and Plant Health Inspection Service. *Animal Welfare Act* 22 (Riverdale: U.S. Department of Agriculture, 1990).
2. NRC (National Research Council). Guide for the Care and Use of Laboratory Animals (Washington, National Academy Press, 1996).
3. Beynen, A.C., Gärtner, K. & van Zutphen, L.F.M. Standardization of animal experimentation In *Principles of Laboratory Animal Science* (eds., van Zutphen, L.F.M., Baumans, V. & Beynen, A.C.) 103–110 (Elsevier, Amsterdam, 2003).
4. Festing, M.F.W. Good experimental design and statistics can save animals, but how can it be promoted? *Altern. Lab. Anim.* **32**, 133–135 (2004).
5. Festing, M.F.W. Refinement and reduction through the control of variation. *Altern. Lab. Anim.* **32**, 259–263 (2004).
6. Quinn, G.P. & Keough, M.J. Hypothesis testing. In *Experimental Design and Data Analysis for Biologists* (eds., Quinn, G.P. & Keough, M.J.) 32–57 (Cambridge University Press, 2002).
7. Crabbe, J.C., Wahlsten, D. & Dudek, B.C. Genetics of mouse behaviour: interactions with laboratory environment. *Science* **284**, 1670–1672 (1999).
8. Wahlsten, D. Standardizing tests of mouse behavior: reasons, recommendations, and reality. *Physiol. Behav.* **73**, 695–705 (2001).
9. Würbel, H. Behavioral phenotyping enhanced—beyond (environmental) standardization. *Genes Brain Behav.* **1**, 3–8 (2002).
10. Campbell, D.T. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* **54**, 297–312 (1957).
11. Lehner, P.N. Handbook of Ethological Methods, 2nd edn (Cambridge University Press, Cambridge, 1996).
12. Guala, F. Experimental localism and external validity. *Philos. Sci.* **70**, 1195–1205 (2003).
13. Van der Staay, F.J. Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. *Brain Res. Brain Res. Rev.* **52**, 131–159 (2006).
14. Crestani, F., Martin, J.R., Möhler, H. & Rudolph, U. Resolving differences in GABA$_A$ receptor mutant mouse studies. *Nat. Neurosci.* **3**, 1059 (2000).
15. Wahlsten, D., *et al.* Different data from different labs: lessons from studies of gene-environment interaction. *J. Neurobiol.* **54**, 283–311 (2003).
16. Wolfer, D.P. *et al.* Laboratory animal welfare: cage enrichment and mouse behaviour. *Nature* **432**, 821–822 (2004).
17. Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G.I. & Golani, I. Genotype-environment interactions in mouse behavior: A way out of the problem. *Proc. Natl. Acad. Sci. USA* **102**, 4619–4624 (2005).
18. Lewejohann, L. *et al.* Environmental bias? Effects of housing conditions, laboratory environment and experimenter on behavioral tests. *Genes Brain Behav.* **5**, 64–72 (2006).
19. Mandillo, S. *et al.* Reliability, robustness and reproducibility in mouse behavioral phenotyping: a cross-laboratory study. *Physiol. Genomics* **34**, 243–255 (2008).
20. Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L. & Mogil, J.S. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. *Neurosci. Biobehav. Rev.* **26**, 907–923 (2002).
21. Valdar, W. *et al.* Genetic and environmental effects on complex traits in mice. *Genetics* **174**, 959–984 (2006).
22. Wahlsten, D., Bachmanov, A., Finn, D.A. & Crabbe, J.C. Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. *Proc. Natl. Acad. Sci. USA* **103**, 16364–16369 (2006).
23. Andrews, N. & File, S.E. Handling history of rats modifies behavioural effects of drugs in the elevated plus-maze test of anxiety. *Eur. J. Pharmacol.* **235**, 109–112 (1993).
24. Rampon, C. *et al.* Enrichment induces structural changes and recovery from nonspatial memory deficits in CA1 NMDAR1-knockout mice. *Nat. Neurosci.* **3**, 238–244 (2000).
25. Cabib, S., Orsini, C., Le Moal, M. & Piazza, P.V. Abolition and reversal of strain differences in behavioural responses to drugs of abuse after brief experience. *Science* **289**, 463–465 (2000).
26. Kas, M.J.H. & Van Ree, J.M. Dissecting complex behaviours in the post-genomic era. *Trends Neurosci.* **27**, 366–369 (2004).
27. Izidio, G.S., Lopes, D.M., Spricigio, L. & Ramos, A. Common variations in the pretest environment influence genotypic comparisons in models of anxiety. *Genes Brain Behav.* **4**, 412–419 (2005).
28. Öbrink, K.J. & Rehbinder, C. Animal definition: a necessity for the validity of animal experiments. *Lab. Anim.* **34**, 121–130 (2000).
29. van der Staay, F.J. & Steckler, T. Behavioural phenotyping of mouse mutants. *Behav. Brain Res.* **125**, 3–12 (2001).
30. van der Staay, F.J. & Steckler, T. The fallacy of behavioral phenotyping without standardisation. *Genes Brain Behav.* **1**, 9–13 (2002).
31. Würbel, H. Behaviour and the standardization fallacy. *Nat. Genet.* **26**, 263 (2000).
32. De Witt, T.J. Scheiner, S.M. *Phenotypic Plasticity. Functional and Conceptual Approaches.* (Oxford University Press, 2004).
33. Latham, N. & Mason, G. From house mouse to mouse house: the behavioural biology of free-living *Mus musculus* and its implications in the laboratory. *Appl. Anim. Behav. Sci.* **86**, 261–289 (2004).
34. Silver, L.M. Mouse genetics: concepts and applications (Oxford University Press, New York, 1995).
35. Silva, A. J., *et al.* Mutant mice and neuroscience: recommendations concerning genetic background. *Neuron* **19**, 755–759 (1997).
36. Beynen, A.C., Festing, M.F.M. & van Montfort, M.A.J. Design of animal experiments In *Principles of Laboratory Animal Science* (eds., van Zutphen, L.F.M., Baumans, V. & Beynen, A.C.) 219–249 (Elsevier, Amsterdam, 2003).
37. Baggerly, K.A. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20**, 777 (2004).
38. Anderson, C.A., Lindsay, J.J. & Bushman, B.J. Research in the psychological laboratory: truth or triviality? *Curr. Dir. Psychol. Sci.* **8**, 3–9 (1999).
39. Vissers, G., Heyne, G., Peters, V. & Guerts, J. The validity of laboratory research in social and behavioral science. *Qual. Quant.* **35**, 129–145 (2001).
40. Boehm, G.W. *et al.* Learning and memory in autoimmune BXSB mouse: Effects of neocorticaol ectopias and environmental enrichment. *Brain Res.* **726**, 11–22 (1996).
41. Chapillon, P., Manneche, C., Belzung, C. & Caston, J. Rearing environmental enrichment in two inbred strain of mice: 1. Effects on emotional reactivity. *Behav. Genet.* **29**, 41–46 (1999).
42. Cudilo, E., Al Naemi, H., Marmorstein, L. & Baldwin, A.L. Knockout mice: is it just genetics? Effect of enriched housing on Fibulin-4$^{+/-}$ mice. *PLoS ONE* **2**, e229, (2007).
43. Hascoet, M., Colombel, M.-C. & Bourin, M. Influence of age on behavioural response in the light/dark paradigm. *Physiol. Behav.* **66**, 567–570 (1999).
44. Imhof, J.T., Coelho, Z.M.I., Schmitt, M.L., Morato, G.S. & Carobrez, A.P. Influence of age and gender on performance of rats in the elevated plus maze apparatus. *Behav. Brain Res.* **56**, 177–180 (1993).
45. Poon, A.M. *et al.* Effect of cage size on ultradian locomotor rhythms of laboratory mice. *Physiol. Behav.* **62**, 1253–1258 (1997).
46. Soffie, M., Hahn, K., Terao, E. & Eclancher, F. Behavioural and glial changes in old rats following environmental enrichment. *Behav. Brain Res.* **101**, 37–49 (1999).
47. Van de Weerd, H.A., Baumans, V., Koolhaas, J.M. & van Zutphen, L.F. Strain specific behavioral response to environmental enrichment in the mouse. *J. Exp. Anim. Sci.* **36**, 117–127 (1994).
48. Russell, W.M.S. & Burch, R.L. The principles of humane experimental technique (Methuen, London, 1959).