Systematic variation improves reproducibility of animal experiments

S Helene Richter, Joseph P Garner, Corinna Auer, Joachim Kunert & Hanno Würbel

Supplementary figures and text:

<table>
<thead>
<tr>
<th>Supplementary Figure 1</th>
<th>Experimental design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Figure 2</td>
<td>Variation of treatment main effects across the four replicate experiments</td>
</tr>
<tr>
<td>Supplementary Figure 3</td>
<td>Between-experiment variation versus within-experiment variation</td>
</tr>
<tr>
<td>Supplementary Figure 4</td>
<td>Effect of standardization and heterogenization on within- and between-experiment variation</td>
</tr>
<tr>
<td>Supplementary Methods</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Information

**Supplementary Figure 1:** Experimental design. 256 female mice of two inbred strains (C57BL/6, BALB/c) were supplied in eight independent batches \((n = 16/\text{strain})\), randomly assigned to cages in groups of four, and allocated to four standardized (STAN) and four heterogenized (HET) replicate experiments conducted in pseudorandom order (1 HET, 1 STAN, 2 STAN, 2 HET, 3 STAN, 4 STAN, 3 HET, 4 HET) over a period of thirteen weeks. Experimental conditions between experiments were varied using eight factors that typically differ between experiments and laboratories and are known to affect a wide variety of potential outcome measures.
Supplementary Figure 2: Variation of treatment main effects across the four replicate experiments. For each experiment, the main effect of ‘strain’ is displayed in terms of the mean F-ratio (+ s.e.m.) across all 36 behavioral measures. F-ratios were calculated separately for each standardized and heterogenized experiment using a general linear model. Based on the $2 \times 2$ factorial design of the heterogenized experiments and to account for microenvironmental differences due to cage position in the rack, each experiment was divided into four blocks and ‘block’ nested within experiment included in the general linear model. Including ‘block’ as a blocking factor in the statistical model allowed us to control for between-block variation, thereby reducing variance in the data and increasing test sensitivity. In contrast to the standardized replicate experiments, main effects of ‘strain’ were highly stable across the heterogenized experiments.
Supplementary Figure 3: Between-experiment variation versus within-experiment variation. To assess the relative weight of between-experiment variation versus within-experiment variation, an F-ratio was calculated that reflects the partitioning of the ‘strain-by-block’ variance between all 16 blocks in the four replicate experiments into variance due to variation between replicate experiments and variance due to variation between blocks within the same experiment. For this, the mean squares for the ‘strain-by-replicate’ interaction term were divided by the mean squares for the ‘strain-by-block’ interaction term. The mean of the resulting F-ratio (+ s.e.m.) across all 36 behavioral measures is shown for both conditions. The difference between the standardized and heterogenized design was tested using the general linear model $y = ‘experimental\ design’ + ‘behavioral\ measure’$. 

a Open-field-test, Path length centre

C57BL/6 mice

BALB/c mice

path length centre (cm), sqrt-transformed

standardized
heterogenized

Experiment

b Novel-object-test, Path length object zone

C57BL/6 mice

BALB/c mice

path length object zone (cm), sqrt-transformed

standardized
heterogenized

Experiment

Supplementary Figure 4: Effect of standardization and heterogenization on within- and between-experiment variation. Data are presented as box-plots showing medians, 25 % and 75 % percentiles, and 5 % and 95 % percentiles (n = 16/strain and experiment). (a and b) Selected measures from two of three behavioral tests showing within- and between-experiment variation across the four replicate experiments (for an example of the third test see Fig. 2): (a) path length within centre zone in the open-field-test, (b) path length within object zone in the novel-object-test.
Supplementary Methods

Animals and housing conditions

The 256 female mice (C57BL/6, \( n = 128 \) and BALB/c, \( n = 128 \)) were obtained from Charles River Laboratories (Sulzfeld, Germany) aged three or nine weeks and were supplied in eight independent batches (\( n = 16/\text{strain} \)). Upon arrival, the mice were randomly assigned to groups of four and housed in conventional polycarbonate cages (according to the experimental design: Type III: 37 cm \( \times \) 21 cm \( \times \) 15 cm or Type IV: 59 cm \( \times \) 38.5 cm \( \times \) 20) with sawdust (GRADE 6, Hellmann Worldwide logistics GmbH & Co.KG, Germany), shelter (MouseHouse, Tecniplast, Italy), and a piece of tissue paper as standard cage equipment. According to the experimental design, some cages contained the following additional enrichment items: pieces of wood (4 \( \times \) 1.6 \( \times \) 1 cm, ABEDD Dominik Mayr KEG, Köflach, Germany), a woolly hamster bed (Woolly Hamsterbett, Trixie, Tarp, Germany), a cardboard roll (5 cm long, diameter: 4 cm), a wooden branch, a cardboard house “Shepherd shack” (Shepherd Speciality Papers, Indulab, Gams, Switzerland), a handful of straw, and a spike of bread wheat (JR Farm, Holzheim-Persenburgheim, Germany). Once a week, cages were cleaned and up to three enrichment items were replaced or added to the enriched cages. All cages were supplied with standard rat and mouse diet (altromin 1324, Lage, Germany) and tap water \textit{ad libitum}.

Depending on position in the rack, cages may differ in local environmental conditions (e.g. temperature, humidity, lighting, disturbance) owing to variation in proximity to ventilation, lights and human traffic. To avoid any position bias, we controlled for cage position in the experimental design. Cages of BALB/c and C57BL/6 mice were balanced with respect to horizontal and vertical position on the rack, and each pair of adjacent cages of C57BL/6
and BALB/c mice was treated as a ‘block’, assuming greater microenvironmental similarity within blocks than between blocks.

Mice were allowed to acclimatize for five weeks before the onset of the experiments. The colony room was maintained at a temperature of 21 ± 1 °C, a relative humidity of 35 ± 5 % and a 12 h light-dark cycle with the lights off at 8.00 p.m.

**Behavioral testing**

Mice were subjected to three common behavioral tests: day 1: free exploration test, day 2: open field test and day 3: novel object test. To monitor health status, mice were weighed prior to testing and directly after.

The apparatus consisted of four adjacent dark grey plastic arenas (50 cm × 50 cm × 40 cm) located in a separate test room, illuminated by four 60 W bulbs adjusted to yield either 60 lx in the centre of the test arenas in the bright light condition or 10 lx in the dim light condition. The test room was maintained at a temperature of 21 ± 1 °C and a relative humidity of 35 ± 5 %. All tests were run during the light phase (0800 - 2000 h). Mice were transported to the test room in their home cages, and tested directly without habituation to the test room. All tests were video-tracked using Noldus EthoVision 3.1 (Noldus Information Technology, Wageningen NL). Testing order of cages was balanced across strain and rack position, and all 4 animals of a cage were tested simultaneously for 10 min on each test. Arenas and objects inside the arenas were cleaned with a 30 % alcohol solution between trials.

**Free exploration test (FET).** Each arena contained a copy of the shelter present in the home cage (MouseHouse, Tecniplast, Italy) that was located in one corner, with the opening directed towards the arena centre. A trial started with the mouse being placed in
front of the house and ended after 10 min. Using EthoVision, various zones were defined (shelter zone, 5 cm zone around the shelter, remaining zone) and the time spent in, the distance traveled in, and the number of entries into each zone were calculated. In addition, the number of fecal boli dropped was counted at the end of each trial. Altogether, seven behavioral measures were included in the final analysis of this test.

**Open field test (OFT).** Mice were placed in the centre of the empty open field arena and videotracked for 10 min. Again, various zones were defined (corners, wall zones of 5, 10 and 18 cm, respectively and the corresponding centre areas) and the time spent in, the distance traveled in, and the number of entries into each zone were calculated. In addition, the total distance traveled during the 10 min session was analyzed and the number of fecal boli dropped was counted at the end of each trial. Altogether, we selected 17 behavioral measures for the analysis.

**Novel object test (NOT).** 24 h after the open field test, the animals were re-exposed for 10 min to the same arena with a transparent glass object (6.0 cm high, diameter 4.5 - 6.0 cm) placed upright in the centre of the arena. In addition to the zones defined for the open field test, two zones surrounding the object at a distance of 5 and 10 cm were defined as exploration zones. The zone defined by the object itself was excluded from the exploration zones to avoid confounding sitting on the object with object exploration. Again, time spent in, distance traveled in, and the number of entries into the various zones were calculated. Object exploration time and frequency were assessed using the time spent in, and the frequency of entering, the exploration zones. In total, eleven behavioral measures were included in the analysis.
Statistical analysis

For the analysis we selected 36 behavioral measures that are commonly assessed in drug-screening or behavioral phenotyping studies. All data were analyzed using General Linear Models (GLM) and post-hoc t-tests. In a GLM, one or more dependent variables ($y_i$), here the behavioral measures, are explained by a number of independent factors ($x_j$). This is done by partitioning the observed variance in the data into variance due to the different independent factors and unexplained variance (residual variance). Variability is measured in terms of the sums of squares (SS), i.e. the squared deviation of each data point from the mean, which is then divided by the degrees of freedom (df) to calculate the ‘mean square’ (MS), i.e. the measure of variability per df. The MS is used to determine the $F$-statistic of the GLM that is given by the ratio of the MS of the independent factor of interest divided by the MS of the residual variance:

$$F = \frac{\text{MS factor}}{\text{MS residual}}.$$

To meet the assumptions of parametric analysis, residuals were examined graphically for homoscedasticity and outliers, and using the Kolmogorov-Smirnov and Shapiro-Wilk tests for normal distribution. When necessary, the raw data were transformed using square-root or logarithmic transformations.

For the analysis we selected 36 behavioral measures that are commonly assessed in drug-screening or behavioral phenotyping studies. In a first step, we determined mean strain differences (mean strain difference = mean C57BL/6 mice - mean BALB/c mean) for all 36 measures to compare variation among the four independent replicate experiments between the standardized and heterogenized design.
Next, we analyzed each experiment separately as if they had been conducted independently by different laboratories. Therefore, we assessed the main effect of ‘strain’ on each of the 36 behavioral measures using a GLM. Based on the $2 \times 2$ factorial design of the heterogenized replicate experiments, and to account for microenvironmental differences due to cage position in the rack, each experiment was divided into four blocks and ‘block’ included as a blocking factor in the GLM:

$$y = \text{strain} + \text{block}$$

Including ‘block’ as a blocking factor in the GLM allowed us to control for between-block variation, thereby reducing variance in the data and increasing test sensitivity.

To explore the difference between the two experimental designs in the variation between replicate experiments further, we analyzed the four replicate experiments (exp) of each experimental design using the GLM

$$y = \text{strain} + \text{exp} + \text{block(exp)} + \text{strain} \times \text{exp} + \text{strain} \times \text{block(exp)}$$

with ‘block’ nested within ‘exp’ to account for the fact that the levels of ‘block’ were different within each replicate experiment. We then compared the resulting F-ratios of the ‘strain-by-experiment’ term between the two experimental designs using a second GLM blocked by behavioral measure:

$$y = \text{experimental design} + \text{behavioral measure}$$
The rationale for using F-ratios for this comparison was twofold, namely (i) that F-ratios are scale invariant, so the different scales of the different test measures became unimportant, and (ii) that F-ratios in a GLM have a discrete null hypothesis (F = 1) which we could test against. The latter is because F-ratios reflect ‘variance components’, so we can think of the true variance of any factor as being = variance due to that factor + residual variance. Thus, in a GLM, if the variance due to the factor under test is 0, then the F-ratio will ideally be 1 (because F-ratio = (variance due to the factor + residual variance)/residual variance).
variance). Therefore, if the average F-ratio of the ‘strain-by-replicate’ interaction term were equal to 1, this would mean that strain differences did not vary between replicate experiments, which would essentially be the same as perfect reproducibility.

To determine how exactly heterogenization reduced between-experiment variation, an additional F-ratio was calculated:

\[ F = \frac{MS \text{ (strain } \times \text{ exp})}{MS \text{ (strain } \times \text{ block})} \]

This F-ratio reflects the partitioning of all the ‘strain-by-block’ variance between all 16 blocks in the four replicate experiments into variance due to between-experiment variation, and variance due to within-experiment variation, i.e. variation between blocks within the same experiment. Because the residual variance is part of both the numerator and the denominator, it cannot affect the resulting F-ratio. Our prediction was that this ratio will be smaller for heterogenized experiments, and ideally = 1. If it were equal to 1 or lower, this would mean that heterogenization generated as much or even more variance between the four blocks within a replicate experiment as exists between replicate experiments. To test this statistically, we used a post-hoc t-test of the null hypothesis that F equals 1.

All statistical tests were conducted using the software package SPSS (version 16.0 for Windows) and differences were considered to be significant at \( P \leq 0.05 \).