PLOS ONE

RESEARCH ARTICLE

Incorporating inter-individual variability in experimental design improves the quality of results of animal experiments

Marloes H. van der Goot[1]*, Marieke Kooij[1], Suzanne Stolte[1], Annemarie Baars[1], Saskia S. Arndt[1‡], Hein A. van Lith[1,2‡]

1 Faculty of Veterinary Medicine, Department Population Health Sciences, Section Animals in Science and Society, Utrecht University, Utrecht, the Netherlands, **2** Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands

‡ These authors are joint senior authors on this work
* m.h.vandergoot@uu.nl

Citation: van der Goot MH, Kooij M, Stolte S, Baars A, Arndt SS, van Lith HA (2021) Incorporating inter-individual variability in experimental design improves the quality of results of animal experiments. PLoS ONE 16(8): e0255521. https://doi.org/10.1371/journal.pone.0255521

Editor: Caitlin Anne Orsini, Univerity of Texas at Austin, UNITED STATES

Received: March 19, 2021

Accepted: July 16, 2021

Published: August 5, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Inter-individual variability in quantitative traits is believed to potentially inflate the quality of results in animal experimentation. Yet, to our knowledge this effect has not been empirically tested. Here we test whether inter-individual variability in emotional response within mouse inbred strains affects the outcome of a pharmacological experiment. Three mouse inbred strains (BALB/c, C57BL/6 and 129S2) were behaviorally characterized through repeated exposure to a mild aversive stimulus (modified Hole Board, five consecutive trials). A multivariate clustering procedure yielded two multidimensional response types which were displayed by individuals of all three strains. We show that systematic incorporation of these individual response types in the design of a pharmacological experiment produces different results from an experimental pool in which this variation was not accounted for. To our knowledge, this is the first study that empirically confirms that inter-individual variability affects the interpretation of behavioral phenotypes and may obscure experimental results in a pharmacological experiment.

1. Introduction

In preclinical experimental animal research, inter-individual variability in phenotypic response is a major source of within-group variability that may negatively affect the power of animal experiments and the reproducibility of their outcomes [1–3]. The exact constitution of inter-individual variability (also referred to as the third component [2] or phenotypic variation [3]) is poorly understood. It is generally accepted however, that the expression of inter-individual variability is the net result of complex interactions between genetic and environmental factors that are partly modulated by epigenetic processes [3,4]. As a result, quantitative traits have even been shown to vary between individuals of the same mouse inbred strain, despite extensive environmental standardization and the use of genetically and microbiologically defined mice of similar age and sex [1,3].

Inter-individual variability however is often not actively accounted for in the design of animal experiments [3]. Traditionally, this type of variation was regarded as part of a larger source of unwanted noise, falling within the same category as other sources of extraneous noise (i.e. measurement error) and unanticipated environmental effects [3]. In contrast to these other sources however, inter-individual variability has been shown to be relatively robust to standardization efforts, distinguishing it from mere noise [2]. Therefore, an increasing body of research focuses on the identification of methods that consider inter-individual differences by systematically incorporating this variation in experimental design and statistical analysis [3,5–12].

The importance of incorporating inter-individual variability in the design of animal experiments has become especially acknowledged in animal models of behavioral dysfunction (e.g., anxiety, depression [5,13,14], post-traumatic stress disorder, [15,16] but also addiction [17] etc.). In humans, the susceptibility between individuals to develop a particular disorder, as well as the response to treatment is known to vary substantially between individuals [13]. Considering this variability may therefore not only improve reproducibility between studies, it may also contribute to an improved understanding of the mechanisms that underlie such inter-individual variability in human patients [5,13]. In this field, several strategies exist to incorporate inter-individual variability as a variable [17,18]. The most prominent strategy considers inter-individual variability between subpopulations within an experimental pool (mostly outbred stocks of rats and mice) by separating experimental animals whose expression of a particular trait (i.e. anxiety, activity) lies on opposing ends of a phenotypic distribution, for example by means of a median, tertiary, quartile split [5,17,18]. Interestingly, this use of selection strategies has indirectly demonstrated how existing subpopulations within an experimental pool may mask the detection of overall group effects, thereby providing examples of how inter-individual variability may confound experimental outcomes [14]. Barbelivien et al. [19], for example identified five sub-populations of Long Evans outbred rats that were characterized by differential levels of baseline impulsive choice behavior. Subsequent administration of d-amphetamine only affected impulse choice behavior when these baseline differences were accounted for, while no effect was found when all animals were pooled in the analysis.

A fundamental principle of good design of animal experiments is that all variables should be controlled except that due to treatment and that all treatment and control groups should be identical, with minimal within-group variability [20,21]. Following these principles, accounting for inter-individual variability in the composition of experimental groups should result in better matched individuals regarding control and treatment, thereby improving the experimental design and the quality of the results. To our knowledge however, the extent to which active incorporation of inter-individual differences in the composition of experimental groups affects the outcome of preclinical animal experiments, has never been directly tested.

In this study, we therefore compared the outcomes of an experimental design in which this inter-individual variability was accounted for, to a design in which this variability was not accounted for, to empirically assess to what extent active incorporation of inter-individual variability indeed alters the interpretation of a standard pharmacological experiment when evaluating the effects of an anxiolytic compound on anxiety-related behavior. To do so we defined the behavioral phenotype of experimental animals on an individual level in a pre-experimental period, and subsequently incorporated this information in the design and statistical analysis of our study.

A priori identification of subgroups within an experimental pool is also common in the aforementioned selection strategies. A disadvantage of these selection strategies however is that in the majority of these studies inter-individual variability is established by means of an artificially predetermined quantile (but see [22,23] for exceptions). Median split strategies may lead to a loss in resolution and power as every value above and below the mean is considered

equal, regardless of its position on the phenotypic distribution [24]. Furthermore, strategies considering upper and lower quantiles only include the outer ends of a phenotypic distribution, rather than the entire study population [25]. These criteria contrast with the generally accepted conceptualization of human psychopathology as a continuum [26], which warrants the exploration of subgroups across the entire study population on the basis of variability in the data itself, rather than a predefined criterion [14].

In the present study we therefore used a data-driven clustering approach to identify meaningful subpopulations within our experimental pool (see below). Furthermore, instead of outbred stocks, we assessed inter-individual variability in mouse inbred strains. As outlined above, inter-individual differences in spontaneous behavior have been repeatedly demonstrated within mouse inbred strains, and have been found to be consistent over time within individuals [27–29]. In fact, inbred strains of mice have been demonstrated to be just as variable as outbred stocks of this species [30].

We expanded on a series of previous studies in which we found that our phenotype of interest, behavioral habituation of anxiety-related responses, may differ within BALB/c, C57BL/6 and various 129 substrains [31,32]. These studies measured behavioral habituation as the change in anxiety and activity related behavior after repeated exposure to a mild aversive stimulus (the modified Hole Board (mHB)). Anxiety is typically regarded as a complex behavioral construct that is expressed by both anxiety-related and activity behaviors [33–35]. Multivariate cluster analyses on the resulting individual response trajectories identified two clusters in which mice grouped together across both anxiety and activity related dimensions: individual response types. These response types were of differential adaptive value, and were displayed by individuals of all three strains.

In the present study, we used the same experimental assay and statistical procedure to first behaviorally characterize BALB/c, C57BL/6 and 129S2 mice on their individual response type (pre-experimental period). Next, we designed a pharmacological experiment in which we systematically incorporated this factor in the composition of our experimental groups. We used a complete randomized block design with four replicates ('mini-experiments or blocks', [36,37]) and systematically incorporated experimenter, besides inbred strain, as a heterogenization factor to improve the generalizability of our results, as suggested by Richter [38].

Previous research showed that anxiolytic compounds may improve behavioral habituation of anxiety responses in the mHB [39]. In this experiment we evaluated the effectiveness of dexmedetomidine as an anxiolytic. In humans this highly selective alpha 2A-adrenergic receptor agonist is reported to exert anxiolytic effects when administered as an analgesic sedative [40]. In mice this compound is used in search for brain mechanisms behind anxiety related behavior, because the alpha 2A-adrenoceptor system is known to play a crucial role in acute neuropsychological stress responses [41].

2. Materials and methods

2.1. Ethical statement

The experimental protocol was approved by the Central Animal Experiments Committee (CCD), the Hague, the Netherlands (CCD approval numbers: AVD1080020172264 and AVD1080020172264-1). The resolution for approval was reached on the basis of the Dutch implementation of EU directive 2010/63/EU (Directive on the Protection of Animals Used for Scientific Purposes). The experiment was conducted according to the Dutch 'Code on Laboratory Animal Care and Welfare'. Furthermore, the present animal study is reported to the best of our abilities according to the revised ARRIVE guidelines (ARRIVE 2.0; https://www.nc3rs.org.uk/revision-arrive-guidelines [42,43]).

2.2. Animals and housing

This study tested naïve males of three mouse inbred strains: BALB/cAnNCrl (hereafter C, $n = 59$, white (albino), strain code = 028), C57BL/6NCrl (B6N, $n = 60$, black, strain code = 027) and 129S2/SvPasCrl (129S2, $n = 60$, agouti, strain code = 287). One additional C mouse died due to reasons unrelated to the study and was not tested.

Furthermore, an additional number of 15 naïve males ($n = 5$/strain) were used to establish the required dose of pharmacological treatment in a pilot study. One B6N mouse died due to reasons unrelated to the pilot study and was not tested. The total number of animals used in the present study amounted to 179 + 14 = 193. The sample size was determined using the software by Lenth (www.stat.uiowa.edu/~rlenth/Power). To eliminate possible effects of estrous cycle on behavioral variables [44], only male mice were used in this study.

Animals were bred by and purchased from Charles River Germany (Sulzfeld, Germany). All mice were 7 weeks old upon arrival (body weight (g), mean ± standard error of the mean (SEM) and range: C, 20.4 ± 0.20 and 13.5–23.4; B6N, 21.0 ± 0.20 and 17.5–24.0; 129S2, 24.1 ± 0.27 and 19.3–28.3). Animals were housed at the Central Laboratory Animal Research Facility of Utrecht University. Testing took place in the same rooms as where the animals were housed, and test equipment was placed in each room prior to arrival of the animals.

Mice were housed individually to reduce aggression and to avoid a potential confounding effect of aggression in (part of) the study population [45,46]. Mice were housed in Macrolon Type II L cages (size: 365 x 207 x 140 mm, floor area 530 cm^2, Techniplast, Milan, Italy) with standard bedding material (autoclaved Aspen Chips, Abedd-Dominik Mayr KEG, Köflach, Austria) and a tissue (KLEENEX® Facial Tissue, Kimberley-Clark Professional BV, Ede, the Netherlands) and a plastic PVC shelter as enrichment (Plexx BV, Elst, the Netherlands). Food (CRM, Expanded, Special Diets Services Witham, UK) and tap water were available *ad libitum*. Upon arrival mice were randomly allocated to one of two laboratory animal housing rooms for a habituation period of 17 days under a reversed 12 h light/12 h dark cycle (lights off at 7:00 AM) with a radio playing constantly as background noise. The number of mice per strain was similar between the two testing rooms. Relative humidity (mean percentage ± SEM) was controlled (room A: 53.5% ± 2.44; room B: 54.8% ± 2.56) with a ventilation rate of 15–20 changes/hour (both rooms) an average room temperature (mean˚C ± SEM) of 21.7˚C ± 0.23 and 21.9˚C ± 0.40 for room A and B, respectively. The mice were handled three times a week during the habituation period by the same two experimenters that conducted the behavioral observations. Handling mice included picking up the mouse at the base of the tail and placed briefly on top of the home cage or on the arm of the experimenter to accustom them to test procedures.

2.3. Modified Hole Board

Mice were tested in the modified Hole Board (mHB), a test for assessment of unconditioned behavior that combines characteristics of an open field, a hole board and a light-dark box [47]. This assay is designed for analyzing a range of anxiety and activity related behaviors and as such is suitable for a complete phenotyping of complex behavioral constructs, such as behavioral habituation of anxiety responses. The mHB has been described extensively elsewhere [48] and is only briefly explained here.

Fig 1 presents a schematic overview of the mHB. The apparatus consists of a grey PVC opaque box (100 x 50 x 50 cm) with a board made of the same material (60 x 20 x 20 cm) functioning as an unprotected area, as it is positioned in the center of box. The board stacks 20 cylinders (diameter 15 mm) in three lines. The area around the board is divided into 10 rectangles (20 x 15 cm) and 2 squares (20 x 20 cm). The periphery was illuminated with red

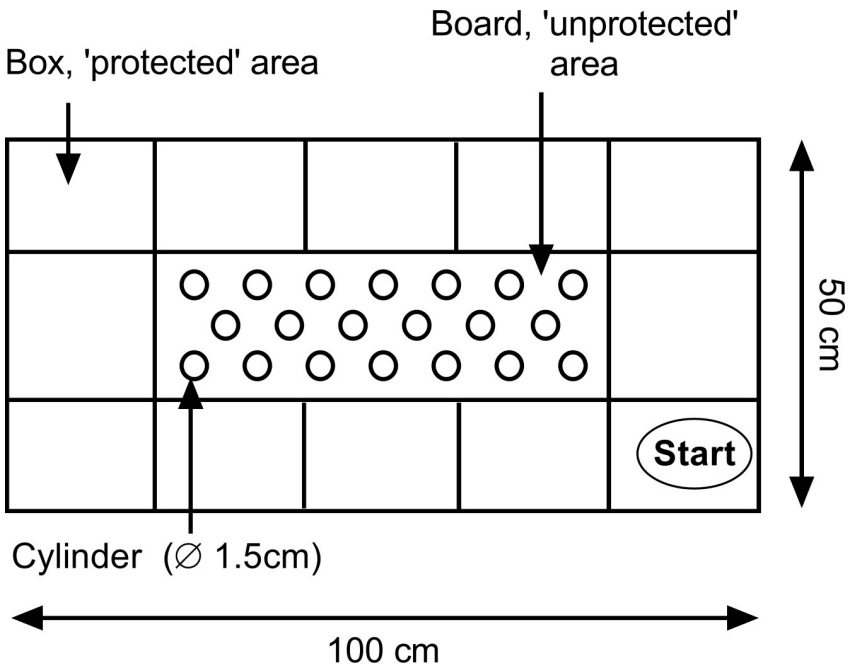

Fig 1. Schematic overview of the modified Hole Board.

https://doi.org/10.1371/journal.pone.0255521.g001

light (1–5 lux) and functioned as the protected area. In contrast, the central board was illuminated by an additional stage light in order to increase the aversive nature of the central (unprotected) area. Light intensity (mean lux) was 147 and 151 lux in room A and B, respectively.

2.4. Experimental protocol

Experimental Phase 1 was used to characterize the mice on their individual response type. In Experimental Phase 2, we designed a pharmacological experiment in which we systematically incorporated this factor (individual response type) in the composition of our experimental groups.

Experimental Phase 1. In order to characterize all 179 mice on their individual response type, the behavior of each individual was repeatedly assessed in the mHB. Each mouse was tested individually for a total of five subsequent trials. Each trial lasted 5 minutes. Behavioral testing occurred between 10 AM and 2 PM, during the active phase of the animals. Data was collected during 6 weeks, with a total number of $n = 30$ ($n = 10$/strain) animals per week. Test order was randomized across strains within each week.

At the start of the first trial, mice were transferred from the home cage to the mHB and always placed in the same corner, facing the central board. During the test, mice were allowed to freely explore the mHB-set up. Between trials mice were picked up by the tail and transported back to their home cage. The mHB was subsequently carefully cleaned with water and a damp towel before the next trial commenced. Behavior was scored live using the software Observer version 12.5 (Noldus Technology, Wageningen, the Netherlands). In addition, trials were recorded on video camera for raw data storage. Behavioral observations were conducted by two trained observers, each of which always tested in the same housing room. For obvious reasons, it was not possible to perform the observations blinded with respect to strain (due to the coat color of the animals, see section 1.2. Animals and Housing). Inter-observer reliability was established at a strong level [49] with an average Cohen's $\kappa = 0.80$ (range 0.71–0.95) over a

percentage agreement of 84.27% (range 76.68–96.35) for frequency scores. For duration scores the inter-observer reliability was established at a strong level, with an average Cohen's $\kappa = 0.88$ (range 0.79–0.95) over a percentage agreement of 92.45% (range 86.20–97.28).

In addition to behavioral observations, circulating corticosterone levels (pCORT) were assessed for each individual mouse at three different time points: one week prior to behavioral testing, directly after the last mHB trial and one week after behavioral testing. These samples were collected with the intention to include pCORT trajectories in our cluster analysis used for classification of the individual response types. However, due to procedural errors during blood sampling and laboratory assay of the plasma there were a substantial number of missing or excluded samples ($n = 30$ samples, of $n = 28$ individuals). To maintain a sufficient sample size and power for the second part of our experiment, individual characterization of mice was therefore only based on their behavioral response.

Experimental Phase 2. In Experimental Phase 2, we systematically incorporated the factor individual response type in the composition of our experimental groups in a pharmacological experiment. A total number of 96 mice were matched to pairs ($n = 48$ pairs, $n = 16$ pairs/strain). The factor individual response type was systematically included by matching half of these pairs on body weight and their individual response type ($n = 24$ pairs, $n = 8$ pairs/strain). This balanced pool represented an experimental design in which individual response type was taken into account.

The remaining half of the pairs were matched on body weight only ($n = 24$ pairs, $n = 8$ pairs/strain). This unbalanced pool mimicked a regular experimental setup in which individual response type was not controlled for. In theory, not accounting for individual response type may result in pairs that share the same response type, or pairs that differ in individual response type. The matched pairs in the unbalanced pool therefore consisted of pairs that shared the same individual response type (62.5%, $n = 13$), and pairs that did not (37.5%, $n = 9$).

Within each pair, one mouse was treated with an anxiolytic, while the other served as control. Treatment and control were assigned randomly within pairs. Treatment consisted of an intra-peritoneal injection (*i.p.*) with dexmedetomidine (Dexdomitor®, 10 μg/kg, 100 μl; Orion Corporation-Orion Pharma, Espoo, Finland). Often used as a sedative-analgesic agent, this pharmacon is a highly selective α_2-adrenergic receptor antagonist that also poses anxiolytic properties [50]. The selected dose was based on a pilot study in which this dose produced behavioral changes but no sedative effect. Control mice of each pair received a saline injection (NaCl, 0.9%, 100 μl, *i.p*). Treatment and control were assigned randomly within pairs.

Pairs of mice were tested over a period of 4 consecutive days, with a weekend in between (thu-fri-mo-tue). The experiment was designed as a complete randomized block design in which each test day was treated as a separate block. The numbers of pairs were maintained equal between test days, between strains and between experimenter. This amounted to 1 balanced, and 1 unbalanced pair (2) per strain (3), per experimenter (2) per block (4), resulting in 48 pairs. Fig 2 presents a schematic representation of the distribution of pairs within a single block (test day), for one experimenter.

At the start of each test day all animals of that test day were weighed to determine the injection volume, 60 minutes prior to start of the first mHB trial. Each mouse was tested individually for a single mHB trial, which lasted for 5 minutes. The experimental procedure of mHB testing was the same as in Experimental Phase 1. Behavioral testing in the mHB again occurred between 10 AM and 2PM. All mice received an intra-peritoneal (*i.p.*) injection (dexmedetomidine or saline) 30 minutes before being placed in the box compartment of the mHB. All injections were given by an experienced technician, who was not involved in the behavioral observations. Pairs of mice were tested after one another, and testing of pairs of the balanced and the unbalanced pool was alternated. Test order of pairs was randomized across strain,

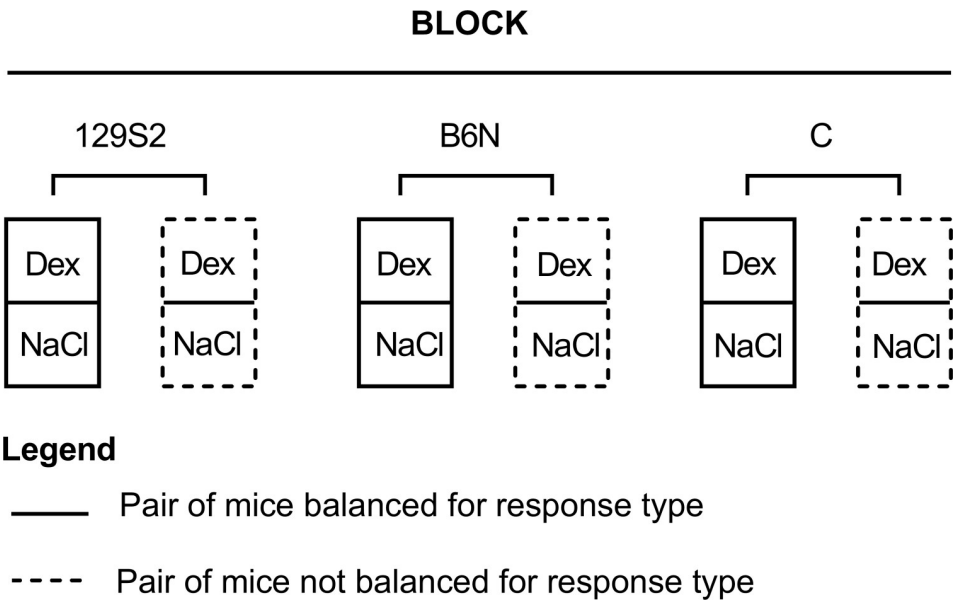

Fig 2. Schematic overview of the 2 (treatment) x 3 (strain) x 2 (experimenter) factorial complete randomized block design used in the pharmacological experiment. Overview represents a single block, for one experimenter. Pairs in the balanced condition were matched on weight and individual response type. Pairs in the unbalanced condition were only matched on weight. The unbalanced condition mimicked a 'regular' animal experiment in which individual response type is not taken into account in assignment to experimental groups. Pairs in the unbalanced condition could therefore consist of one animal from cluster A, and one animal from cluster B, or consist of two animals from the same cluster. Pairs in the balanced condition only consisted of two animals of the same cluster. Within each of 4 blocks (4 testing days) for each experimenter, all pairs were matched within strain, and within experimenter. Dex = treatment with dexmedetomidine (Dexdomitor®, 10 μg/kg, 100 μl, *i.p*); NaCl = control treatment with saline (NaCl, 0.9%, 100 μl, *i.p*).

https://doi.org/10.1371/journal.pone.0255521.g002

experimenter and test day. Behavioral observations were conducted by the same two experimenters, each in the same housing room, as in Experimental Phase 1, using the same ethogram and the same software. These experimenters were blind to treatment, pair and whether the pair was balanced or unbalanced on individual response type.

2.5. Behavioral variables

Experimental Phase 1 and 2. Behavioral patterns of mice were assessed by scoring behaviors listed in Table 1. From these observations, several parameters for avoidance behavior, exploration and locomotion were computed (Table 1).

However, previous studies using the mHB have shown that these separate behaviors scored in this assay can be reliably summarized to underlying behavioral dimensions; avoidance behavior, exploration and locomotion [51–53]. Two previous studies showed that simultaneous clustering of these behavioral dimensions yields distinct response types that are displayed by individuals of C, B6N and 129S2 [31,32].

For each dimension, observed behaviors indicative of that dimension were summarized to integrated behavioral z-scores according to the procedure described in Labots et al. [53] and van der Goot et al. [31]. In short, this entailed that behavioral variables measuring different aspects of the same behavioral dimension were normalized to z-scores, and combined to a single integrated z-score representing that dimension. Combination of z-scores was done by averaging them. For normalization of each separate variable, we used the pooled data (across all

Table 1. Behavioral parameters measured in the modified Hole Board.

Behavioral dimension	Behavioral parameter	Description of mouse behavior
Avoidance behavior	Total number of board entries	Mouse on the central board
	Latency until the first board entry	
	Percentage of total time spent on board	
Exploration	Total number of rearings in the box	Rearing on hind paws in the box
	Latency until the first rearing in the box	
	Total number of rearings on the board	Rearing on hind paws on the board
	Latency until the first rearing on the board	
	Total number of hole explorations	Exploration of a cylinder (hole) on the board
	Latency until the first hole exploration	
	Total number of hole visits	Nose-poking into a cylinder (hole) on the board
	Latency until the first hole visit	
Locomotion	Total number of line crossings	Line crossing with all its paws in the box
	Latency until the first line crossing	

https://doi.org/10.1371/journal.pone.0255521.t001

strains) as a reference group, as suggested by Labots et al. [53]. An overview of all included variables per dimension is listed in S1 Table.

2.6. Statistics

All analyses were conducted with R version 4.0.0 in R-Studio [54]. Linear mixed models (LMMs) were run using the package 'nlme' [55]. The package 'kml3d' was used for running a multivariate cluster analysis on longitudinal response trajectories [56]. All Figures were created with GraphPad Prism (GraphPad Prism version 7.04 for Windows, Graphpad Software, La Jolla, California USA, www.graphpad.com).

Experimental Phase 1. The total number of individuals included for behavioral analysis per strain was C (n = 59), B6N (n = 60) and 129S2 (n = 60). The behavioral variables obtained in this phase were used to characterize mice on their individual response type (Table 1). Additional analyses assessing between strain differences in behavioral scores, as well as the collected pCORT data were not reported in the present paper because they fall beyond the scope of the manuscript. These results will therefore be reported elsewhere.

The procedure described in section 4.5 yielded three trajectories of integrated residual z-scores for each individual mouse, one trajectory per behavioral dimension: avoidance behavior, exploration and locomotion. These three trajectories were fit with LMMs to control for potentially confounding factors. The resulting standardized Pearson residuals could then be used for a clustering procedure.

For each behavioral dimension, potentially confounding variables were controlled for by including strain and experimenter as fixed factors, and individual mouse, test group and test order as random factors in the model. The variable 'trial' was intentionally left out of the model because we wanted to maintain this information in the residuals so that we could assess habituation of individual mice over time. Models were run with an autoregressive correlation structure for continuous time covariates (corCAR1) from the 'nlme' package.

Model assumptions were assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots [57,58]. The dimension avoidance behavior was logarithmically transformed to achieve normality of the residuals. A square root transformation

was applied on exploration and locomotion was rank transformed. Heteroscedasticity was avoided using the 'varIdent' variance structure function from the 'nlme' package, allowing different residual spread for each level of the categorical variables in our model [58]. The dimensions avoidance behavior, exploration and locomotion included a variance function for 'Strain'.

The resulting standardized Pearson residual integrated z-score trajectories were subsequently analyzed with a multivariate k-means clustering procedure for longitudinal data, *kml3d* [56]. The settings and rationale for using this method have been described in detail in [31]. Furthermore, the settings used in the present manuscript are identical to a previous study [32].

As described, three response trajectories were included for each individual mouse: avoidance behavior, exploration and locomotion. These were clustered simultaneously to assess the occurrence of homogenous subgroups of mice that shared similar responses (between them) on all three behavioral dimensions. Prior to analysis, the gap statistic was applied to evaluate whether the data was perhaps best represented by a single cluster using the package 'cluster' [59]. This was not the case. The gap statistic compares the within-cluster sum-of-squares to a null reference distribution of the data, which is then equivalent to a single cluster [60], and as such gives an indication of whether it is appropriate to partition the data into clusters. The cluster analysis compiled 1000 iterations for each k clusters between 2 and 6, resulting in 5000 cluster solutions.

The optimal number of clusters was selected using the approach of Clustering Validity Indices (CVI's) as suggested by Kryszczuk and Hurley [61] and adjusted by Wahl et al. [62]. All details of this procedure are described in [31]. After obtaining the optimal clustering solution, we applied a bootstrapping procedure to determine the stability of the identified clusters. 200 random samples (of $n = 179$) were drawn from the original data with replacement, meaning that a particular individual could occur multiple times in one sample. If our clusters were stable, applying the multivariate clustering procedure in these 200 random samples should reveal similar cluster structures [63]. Similarity in cluster composition between the original analysis and the bootstrapping, samples was established by the Jaccard Index. For each individual mouse, the number of times (out of 200 bootstrap samples) it retained its original cluster was determined using the following formula: *number of times in the same cluster/total number of bootstrapping samples*. The individual similarity indices were subsequently averaged across mice to determine the overall Jaccard similarity index for each cluster.

Finally, to characterize the resulting clusters, LMMs analyzed the differences in integrated behavioral z-scores across trials between clusters on each behavioral dimension. Model assumptions and settings were identical to the settings described above. Cluster and trial were included as fixed predictors, as well as their interaction. Individual mouse (ID) and slope (trial nested in ID) were included as random factors. The integrated behavioral z-score for locomotion was rank transformed to improve the residual distribution. A variance function ('varIdent') was applied for Cluster in the model for avoidance behavior, to avoid heteroscedasticity. The model for exploration included a variance function on trial, while locomotion included a variance function for the different trials within Cluster.

Main and interaction effects from all LMMs were derived using F-tests with corresponding P value ($P < 0.05$). Statistical significance of random effects was computed by means of likelihood ratio tests and reported as Chi Square values. Pairwise comparisons were conducted using the package 'emmeans' [64] to follow up on main or interaction effects. To reduce the probability of a Type I error due to multiple comparisons, the α was adjusted using a Dunn-Šidák correction in all *post hoc* tests [65]. S2 Table lists an overview of all corrected α-values used in this manuscript. All *post hoc* tests were summarized as beta-estimates and their

corresponding standard error, *t* statistic and *P* values. Effect sizes for *post hoc* tests were reported as Cohen's *d*, and obtained via the package 'emmeans' [64]. The guidelines provided by Wahlsten [66] were used to interpret the absolute values of Cohen's *d* (|d|). This extensive review of various phenotypes suggested the following interpretation of effects for neurobehavioral mouse studies: small effect, $|d| < 0.5$; medium effect, $0.5 < |d| < 1.0$; large effect, $1.0 < |d| < 1.5$; very large effect, $|d| > 1.5$.

Experimental Phase 2. The results of phase 2 were first analyzed on the combined data of the balanced and the unbalanced pool ($n = 96$ individuals, 48 pairs, $n = 16$ pairs/strain). Whether pairs were balanced or unbalanced on individual response type was incorporated in the factor 'pool' (2 levels, balanced/unbalanced). This dataset, with considerable sample size and power, allowed us to analyze treatment and strain effects while asking whether incorporating the factor 'pool' (accounting for individual response type versus not accounting for this variation) in our analyses would explain part of the variance in our model. For each behavioral dimension, generalized linear models (GLMs) analyzed the effect of dexmedetomidine between strains on integrated behavioral z-scores. Treatment, strain, pool and experimenter were included as fixed factors, as well as all interactions. The factor 'block' (representing test day, see *section 1.4 Experimental protocol*) was included as a random factor without any interactions (as suggested by Festing [67]).

In addition, a second series of GLMs analyzed treatment and strain effects separately for the balanced pool (pairs of mice that were balanced on individual response type, $n = 48$, 24 pairs, $n = 8$/strain), and the unbalanced pool (pairs of mice that were not balanced on response type, $n = 48$, 24 pairs, $n = 8$/strain). For each pool, GLMs analyzed the effect of dexmedetomidine on integrated behavioral z-scores. For each behavioral dimension, treatment, strain and experimenter were included as fixed factors, as well as all interactions. The factor block was included as a random factor, without interactions. Model assumptions were assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots [57,58].

For all analyses in phase 2, the integrated behavioral z-score for exploration was logarithmically transformed and that of locomotion was rank transformed to improve the residual distribution. In addition, Cooks distance identified locomotion scores of 4 mice as influential–these were 2 individuals that had not displayed any line crossing, resulting in the maximum score for latency to first line crossing (300 seconds) and 2 individuals with a latency to first line crossing > 180 seconds. These observations were retained for analysis.

Main and interaction effects from all GLMs were derived using *F*-tests with corresponding *P* value ($P < 0.05$). Effect sizes for the GLMs are reported as partial eta squared (η_p^2) with 95% CI, using the following cut-off limits: small effect, $\eta_p^2 \leq 0.03$; medium/moderate effect, $0.03 < \eta_p^2 < 0.10$; large effect, $0.10 \leq \eta_p^2 < 0.20$; very large effect, $\eta_p^2 \geq 0.20$ [68]. Pairwise comparisons were conducted using the package 'emmeans' [64] to follow up on main or interaction effects. To reduce the probability of a Type I error due to multiple comparisons, the α was adjusted using a Dunn-Šidák correction in all *post hoc* tests [65], see S2 Table. All *post hoc* tests were summarized as beta-estimates and their corresponding standard error, *z* statistic and *P* values. Effect sizes for *post hoc* tests were reported as Cohen's *d*, and obtained via the package 'emmeans' [64]. The guidelines provided by Wahlsten [66] were again used to interpret the absolute values of Cohen's *d* (|d|), see *"Experimental Phase 1"*.

3. Results

3.1. Cluster analysis

The optimal partitioning of the data yielded two clusters, A and B. The table in Fig 3A presents cluster size and distribution of strains across clusters. The majority of individuals (57.5%,

a.

Cluster size (*n*) and proportion of total *n* per cluster				
	Cluster A		Cluster B	
n total = 179	*n* =103 *(57.5%)*		*n* = 76 *(42.5%)*	
Distribution of strains *within* clusters				
	Cluster A		Cluster B	
(sub-) Strain	*n*	%	*n*	%
C	7	6.8	52	68.4
B6N	42	40.8	18	23.7
129S2	54	52.4	6	7.9

b.

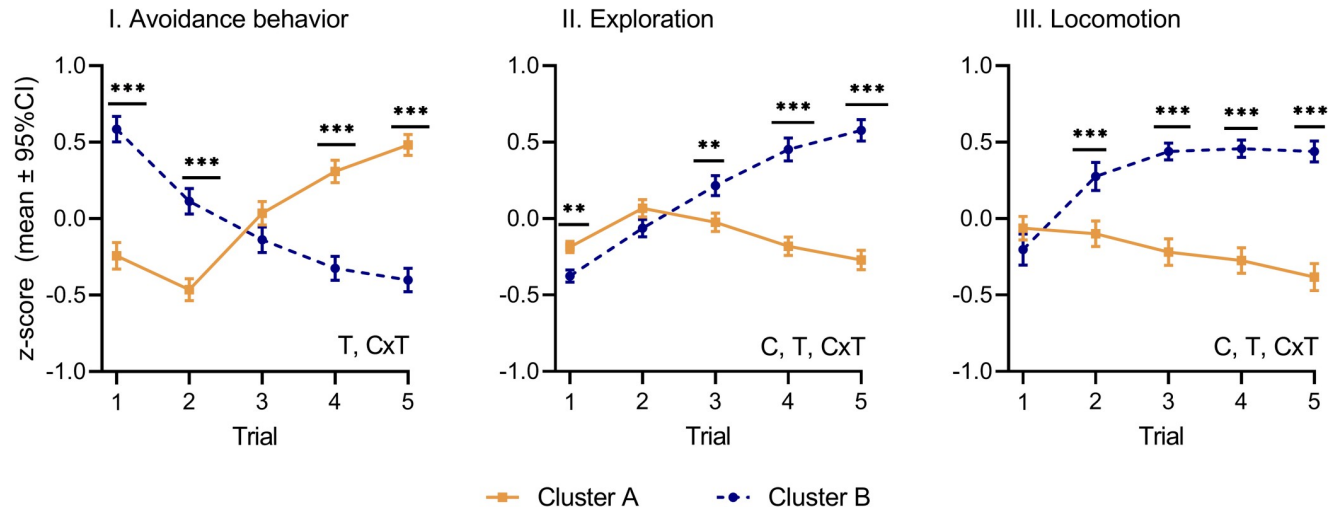

Fig 3. **(a)** Results cluster analysis. Top row: Cluster size and proportion of total population for each cluster (total number of mice, *n* = 179). Bottom rows: Distribution of strains (*n* and proportion) within each cluster. **(b)** Behavioral response trajectories of clusters on avoidance behavior, exploration and locomotion (cluster A, orange; cluster B, blue). Behavior expressed as integrated behavioral z-scores. Results are presented as means with 95% CI. Effects were significant in LMMs at $P < 0.05$. C: Significant main effect of cluster; T: Significant main effect of trial; C x T: Significant interaction between cluster and trial. Significant differences in *post hoc* contrasts between clusters on trials 1 and 5 (adjusted $\alpha = 0.025321$) are indicated by ** = $0.00050 \leq P < 0.00501$, *** = $P < 0.00050$. Significant differences in *post hoc* contrasts between clusters on trials 2, 3 and 4 ($\alpha = 0.05$) are indicated by ** = $0.001 \leq P < 0.01$, *** = $P < 0.001$. The raw scores (mean integrated behavioral z-score ± 95% CI) per trial, per cluster, for each behavioral dimension are listed in S3 Table.

https://doi.org/10.1371/journal.pone.0255521.g003

n = 103) grouped together in cluster A while the remaining mice formed cluster B (42.5%, *n* = 76). Each cluster consisted of mice from all three strains. The majority of C mice (88.1%, *n* = 52) grouped together in cluster B while the majority of 129S2 (90%, *n* = 54) and the majority of B6N (70%, *n* = 42) grouped together in cluster A. The clusters displayed differential patterns of behavior on all three behavioral dimensions, as indicated by significant interactions between clusters and trial (Avoidance behavior, trial effect: $F_{(4,708)} = 13.17$, $P < 0.0001$; interaction cluster x trial: $F_{(4,708)} = 46.58$, $P < 0.0001$; Exploration, cluster effect: $F_{(1,177)} = 12.12$, $P = 0.0006$; trial effect: $F_{(4,708)} = 37.63$, $P < 0.0001$; interaction cluster x trial: $F_{(4,708)} = 44.97$, $P < 0.0001$; Locomotion, cluster effect: $F_{(1,177)} = 78.43$, $P < 0.0001$; trial effect: $F_{(4,708)} = 4.85$, $P = 0.0007$; interaction cluster x trial: $F_{(4,708)} = 19.64$, $P < 0.0001$; Fig 3B).

Post hoc comparisons (adjusted $\alpha = 0.025320$) showed that mice in cluster A increased avoidance behavior between the first and the last trial (-0.726 ± 0.96, $t_{(708)} = -7.593$, $P < 0.0001$, *large* effect size, $d = -1.015$, 95%CI [-1.283, -0.747]). At the same time, locomotion (rank transformed) decreased (98.06 ± 23.1, $t_{(708)} = 4.250$, $P < 0.0001$, *small* effect size, $d = 0.498$, 95%CI [0.267, 0.730]), while exploration remained stable across trials (0.085 ± 0.05, $t_{(708)} = 1.583$, not significant, S4-I Table).

Mice in cluster B displayed the opposite pattern and decreased avoidance behavior between the first and the last trial (0.985 ± 0.96, $t_{(708)}$ = 10.288, $P < 0.0001$, *large* effect size, d = 1.377, 95%CI [1.105, 1.650]), while exploration and locomotion increased (exploration, -0.953 ± 0.06, $t_{(708)}$ = -15.274, $P < 0.0001$, *very large* effect size, d = -3.588, 95%CI [-4.085, -3.090]; locomotion (rank transformed), -252.68 ± 32.6, $t_{(708)}$ = -7.740, $P < 0.0001$, *large* effect size, d = -1.285, 95%CI [-1.618, -0.952]), see S4-I Table.

The trajectories of avoidance behavior were significantly higher in cluster B on the first two trials (trial 1, $P < 0.0001$, *large* effect size, d = -1.158, 95%CI [-1.488, -0.827]; trial 2, $P < 0.0001$, *medium* effect size, d = -0.808, 95%CI [-1.127, -0.489]) and lower than cluster A on trials 4 and 5 (trial 4, *medium* effect size, $P < 0.0001$, d = 0.887, 95%CI [0.565, 1.208]; trial 5, *large* effect size, $P < 0.0001$, d = 1.235, 95%CI [0.901, 1.569], Fig 3B-I and S4-II Table). Furthermore, exploration in cluster B was lower on trial 1 (P = 0.0011, *medium* effect size, d = 0.715, 95%CI [0.283, 1.147]) and higher on the last three trials (trial 3, P = 0.0042, *medium* effect size, d = -0.903, 95%CI [-1.525, -0.281]; trial 4, $P < 0.0001$, *very large* effect size, d = -2.384, 95%CI [-3.135, -1.635]; trial 5, $P < 0.0001$, *very large* effect size, d = -3.192, 95%CI [-4.009, -2.375], Fig 3B-II and S4-II Table). Locomotion was significantly higher in cluster B on all trials except trial 1 (trial 2, $P < 0.0001$, *medium* effect size, d = -0.774, 95%CI [-1.140, -0.406]; trial 3, $P < 0.0001$, *large* effect size, d = -1.157, 95%CI [-1.510, -0.798]; trial 4, $P < 0.0001$, *large* effect size, d = -1.384, 95%CI [-1.730, -1.033]; trial 5, $P < 0.0001$, *very large* effect size, d = -1.562, 95%CI [-1.910, -1.208], Fig 3B-III and S4-II Table).

3.1.1 Cluster stability. Cluster stability was assessed with a bootstrapping procedure in which 200 random samples (of n = 179) were drawn from the original data with replacement. The trajectories of the original cluster and the average of all 200 cluster analyses were highly similar (Fig 4A).

The average Jaccard Index was 0.92 for cluster A (Fig 4B), meaning that on average, mice that fell in cluster A also did so in 92% of the bootstrap samples. The average Jaccard Index for cluster B was equally high (0.95, Fig 4B). The originally obtained clusters A and B were thus rendered stable, and representative for this dataset.

3.2. Results pharmacological experiment

The cluster analysis revealed two differential response types, which were displayed by individuals of all three strains. We next explored whether incorporating these individual response types in the design of a standard pharmacological experiment would affect the results in comparison to an experiment in which this variation was not controlled for. A 2 (treatment) x 3 (strain) x 2 (experimenter) factorial complete randomized block design was used to test the effect of dexmedetomidine on behavior in the mHB.

3.2.1 Incorporating individual variation as a discriminating factor in analysis. The results were first analyzed on the total population (n = 96, 48 pairs, n = 16 pairs/strain), the combined data of the unbalanced and the balanced pool. Generalized linear models (GLMs) analyzed the effect of dexmedetomidine on behavior using a 2 (treatment) x 3 (strain) x 2 (pool) x 2 (experimenter) factorial design, including all interactions. The factor 'block' (test day, n = 4) was included as a random factor without any interactions [67].

Treatment with dexmedetomidine primarily reduced activity related behavior, compared to a control injection with saline. Treated mice displayed less exploration ($F_{(1,68)}$ = 7.36, P = 0.0085, *medium* effect size, η_P^2 = 0.097, 95% CI [0.008, 0.252]) and less locomotion than controls ($F_{(1,68)}$ = 9.75, P = 0.0027, *large* effect size, η_P^2 = 0.124, 95% CI [0.016, 0.279]; Fig 5A-II and 5A-III) and S5 Table). The effect of dexmedetomidine on anxiety related behavior was less pronounced, but there was a suggestion of average higher levels of avoidance behavior

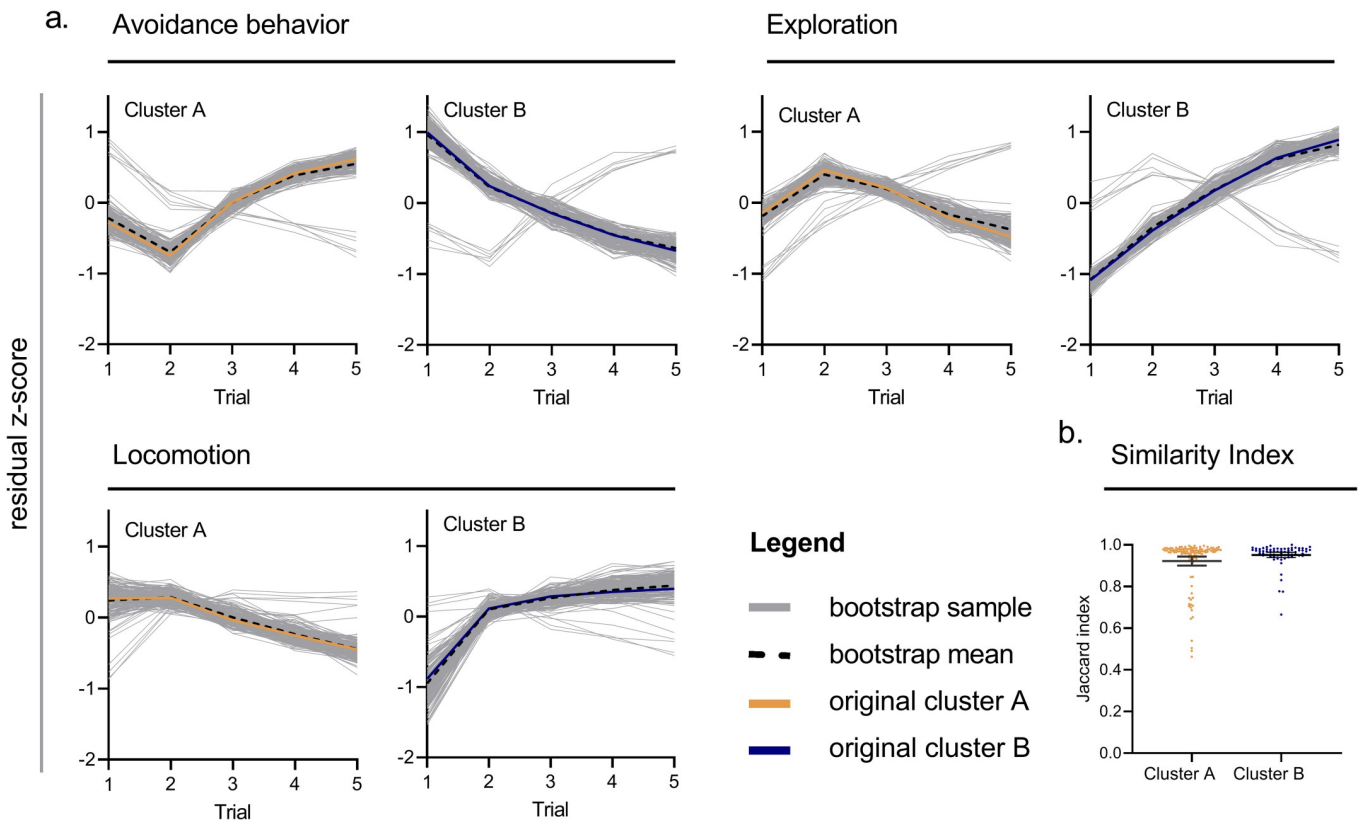

Fig 4. **(a)** Results of the bootstrapping procedure for each cluster, on each behavioral dimension. Results are presented as mean residual z-scores, and depict the trajectory of the original cluster (cluster A, orange; cluster B, blue) in relation to the average of all 200 bootstrap samples (black dashed line), against all 200 trajectories of the bootstrapping procedure (grey). **(b)** Distribution of individual Jaccard Indices in clusters A and B (cluster A, orange dots; cluster B, blue dots). Average Jaccard Index per cluster indicated by mean with 95% CI (black).

https://doi.org/10.1371/journal.pone.0255521.g004

in treated animals ($F_{(1,68)}$ = 3.70, P = 0.0586, *medium* effect size, η_P^2 = 0.052, 95% CI [0.000, 0.185]; Fig 5A-I and S5 Table).

Strain also differed significantly in exploration ($F_{(2,68)}$ = 13.38, P = 0.0000) and locomotion ($F_{(2,68)}$ = 29.23, $P < 0.0001$), both with very large effect sizes (respectively η_P^2 = 0.282, 95%CI [0.104, 0.423]; η_P^2 = 0.493, 95%CI [0.319, 0.616]; Fig 5B-II and 5B-III and S5 Table).

Post hoc comparisons (adjusted α = 0.02532) showed that exploration was highest in B6N compared to C and 129S2 (C, P = 0.0002, *medium* effect size, d = -0.928, 95%CI [-1.499, -0.408]; 129S2, $P < 0.0001$, *large* effect size, d = -1.327, 95%CI [-1.891, -0.763]; Fig 5B-II and S6A Table). Furthermore, locomotion differed between all three strains, with higher levels of locomotion in B6N compared to C and 129S2 (C, $P < 0.0001$, *large* effect size, d = -1.240, 95% CI [-1.780, -0.700]; 129S2, $P < 0.0001$, *very large* effect size, d = -2.000, 95%CI [-2.620, -1.381]) and higher locomotion in C than in 129S2 (P = 0.0028, *medium* effect size, d = -0.760, 95%CI [-1.270, -0.246], Fig 5B-III and S6A Table).

Avoidance behavior did not significantly differ between strains ($F_{(2,68)}$ = 2.41, P = 0.0972, medium effect size, η_P^2 = 0.068, 95% CI [0.000, 0.227]). A significant effect of pool however, demonstrated that avoidance behavior was significantly lower in pairs that were matched on response type, than in pairs that were not matched on response type ($F_{(1,68)}$ = 5.37, P = 0.0235, *medium* effect size, η_P^2 = 0.074, 95%CI [0.000, 0.209]; Fig 5C-I and S5 Table). A suggestion of a similar effect (in reversed direction) was found for exploration ($F_{(1,68)}$ = 3.54, P = 0.0643, medium effect size, η_P^2 = 0.049, 95%CI [0.000, 0.175]; Fig 5C-II and S5 Table).

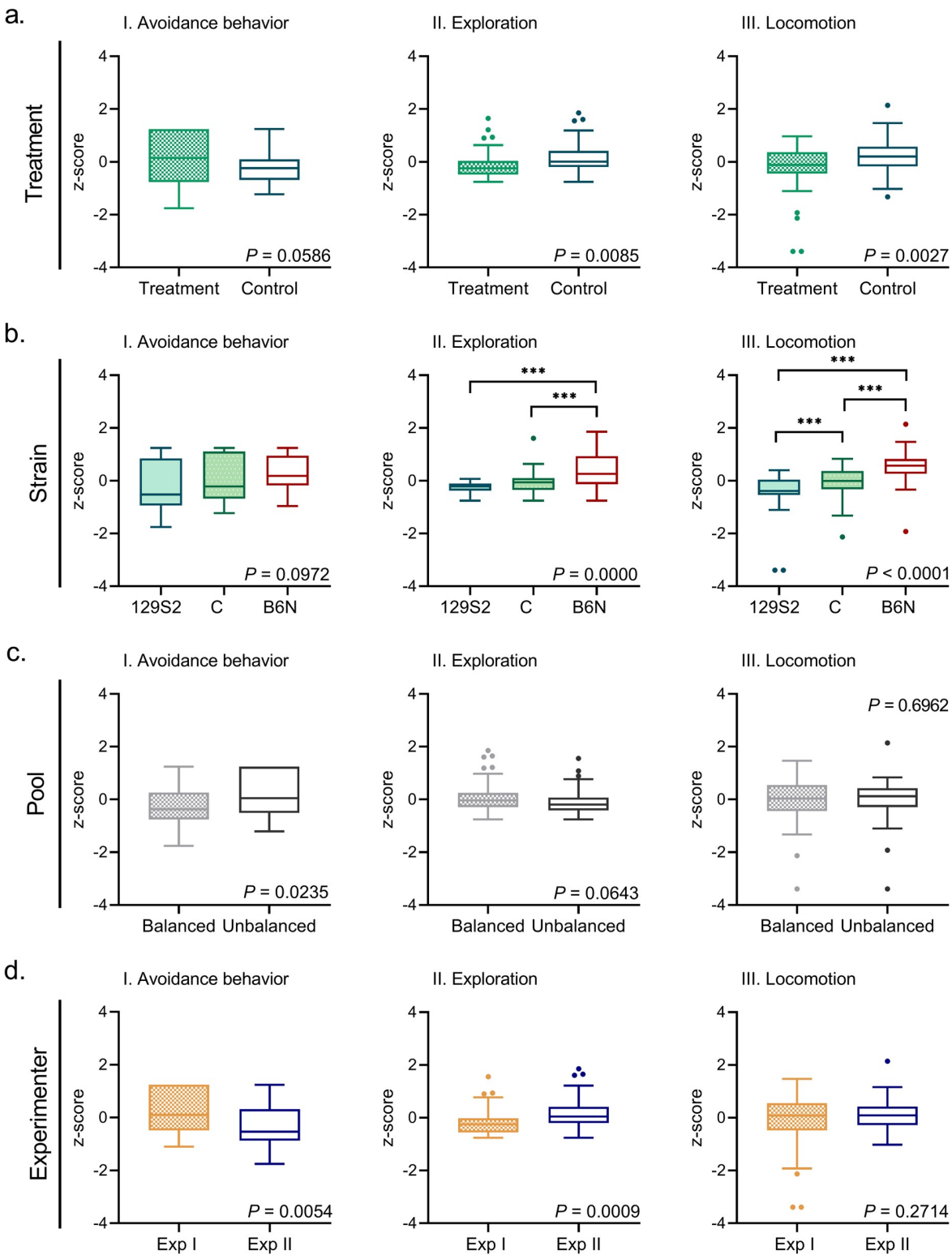

Fig 5. (a-d) Avoidance behavior, exploration and locomotion of mice in a single 5-minute mHB trial. Behavior in all graphs expressed as integrated behavioral z-score. Results are presented as boxplots (median, upper and lower quartiles) with Tukey whiskers. Individual points = outside the upper/lower quartile* 1.5 inter-quartile range. Effects were significant in GLMs at $P < 0.05$. Significant differences in *post hoc* contrasts between strains (adjusted $\alpha = 0.02532$) are indicated by *** when $P < 0.00050$. The raw integrated z-scores (mean ± 95% CI) of each group depicted in this figure are presented in S5 Table.

https://doi.org/10.1371/journal.pone.0255521.g005

The results also revealed experimenter effects for avoidance behavior and exploration: Experimenter I scored higher levels of avoidance behavior and lower levels of exploration than Experimenter II (Avoidance behavior, $F_{(1,68)} = 8.26$, $P = 0.0054$, *large* effect size, $\eta_P^2 = 0.106$, 95%CI [0.010, 0.254]; Exploration, $F_{(1,68)} = 11.95$, $P = 0.0009$, *large* effect size, $\eta_P^2 = 0.150$, 95% CI [0.027, 0.301]; Fig 5D-I and 5D-II and S5 Table).

All in all, these results indicate that behavioral scores may differ between an experimental pool in which individual differences are accounted for, and a pool in which this variation is not incorporated (such as for avoidance behavior). The absence of any significant interactions however suggest that this effect of pool did not interfere directly with treatment or strain effects.

3.2.2 Comparison between balanced and unbalanced pool. Next, we analyzed the balanced and the unbalanced pool separately and compared the results. Aside for controlling for individual response type, we considered the two pools directly comparable with respect to other factors such as experimenter, strain, treatment etcetera. Any difference in results between these two pools of mice was therefore attributed to the fact that we matched our pairs on individual response type in one pool (balanced) and not in the other (unbalanced).

Furthermore, following the principle of good experimental design described in the introduction, matching our pairs on individual response types in the balanced pool should make the treatment and control groups more similar, thereby improving the quality of our results. Any differences in results between the balanced and the unbalanced pool were thus interpreted in favor of the balanced data.

For each pool, GLMs analyzed the effect of dexmedetomidine on behavior using a 2 (treatment) x 3 (strain) x 2 (experimenter) factorial design, including all interactions. The factor 'block' (test day, $n = 4$) was included as a random factor without any interactions [67].

Separate analyses of the balanced and the unbalanced pool indeed yielded different results, especially with respect to exploration and locomotion. In the unbalanced pool, treatment effects on exploration differed between strains (strain effect: $F_{(2,32)} = 9.17$, $P = 0.0007$, *very large* effect size, $\eta_P^2 = 0.364$, 95% CI [0.088, 0.538]; treatment effect: $F_{(1,32)} = 9.13$, $P = 0.0049$, *very large* effect size, $\eta_P^2 = 0.222$, 95%CI [0.023, 0.433]; interaction strain x treatment: $F_{(2,32)} = 4.98$, $P = 0.0131$, *very large* effect size, $\eta_P^2 = 0.238$, 95%CI [0.000, 0.428]; Fig 6II-1 and S7 Table).

Post hoc comparisons between strains in each condition (adjusted $\alpha = 0.01274$) revealed that exploration in the unbalanced pool only differed between strains in the control groups: saline injected B6N displayed more exploration than saline injected C and 129S2 (C, $P = 0.0002$; *very large* effect size, $d = -1.901$, 95%CI [-2.997, -0.805]; 129S2, $P < 0.0001$, *very large* effect size, $d = -2.531$, 95%CI [-3.691, -1.371]; Fig 6II-1 and S6C Table).

Furthermore, *post hoc* comparisons between conditions within strain (adjusted $\alpha = 0.016952$) showed that saline injected B6N displayed more exploration than their counterparts treated with dexmedetomidine ($P < 0.0001$, *very large* effect size, $d = 2.105$, 95%CI [0.997, 3.213]; Fig 6II-1 and S6C Table).

Interestingly, this treatment effect disappeared when analyzing the balanced pool. Exploration still differed between strains (strain effect: $F_{(2,32)} = 8.24$, $P = 0.0013$, *very large* effect size, $\eta_P^2 = 0.340$, 95%CI [0.070, 0.518]) but exploration was now higher in B6N than C and 129S2 regardless of treatment, as opposed to only in the saline condition (C, $P = 0.0031$; *large* effect size, $d = -1.053$, 95%CI [-1.800, -0.310]; 129S2, $P = 0.0001$, *very large* effect size, $d = -1.719$, 95%CI [-2.690, -0.743]; Fig 6II-2 and S6B Table). Also, the treatment effect within B6N disappeared in the balanced condition (Fig 6II-2).

A similar shift of the effect of treatment was found for locomotion. In the unbalanced pool, locomotion was significantly higher in controls compared to treated mice across strains ($F_{(1,32)}$

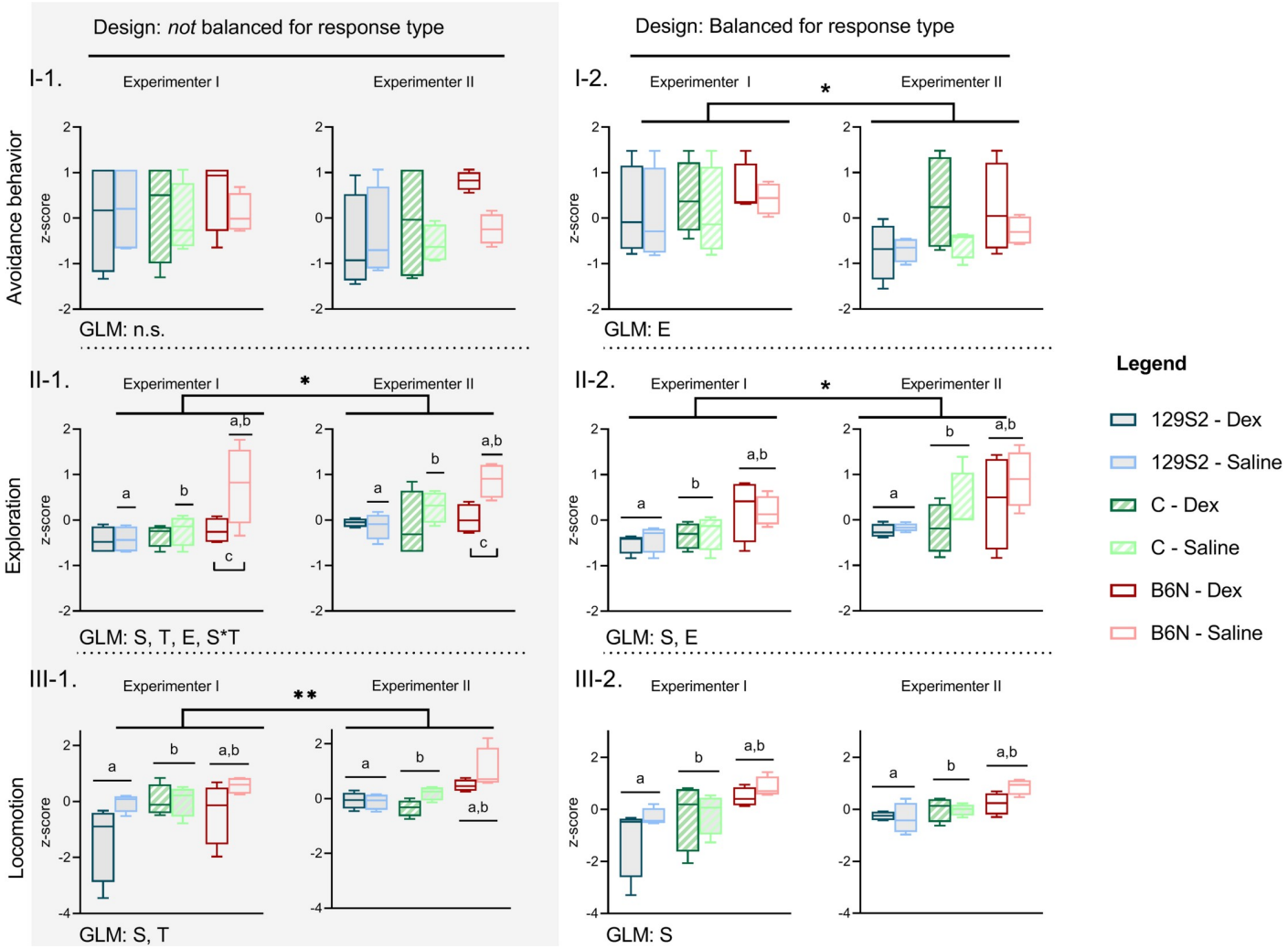

Fig 6. Behavioral scores of a pool of mice not balanced for individual response type (left), and a pool balanced for individual response type (right). **(I-III)** Results expressed as integrated behavioral z-scores and presented as boxplots (median, upper and lower quartiles) with Tukey whiskers. Effects significant in GLMs at $P < 0.05$, where $^{*} = 0.01 \leq P < 0.05$, $^{**} = 0.001 \leq P < 0.01$. T: Significant effect treatment; S: Significant effect strain; S*T: Significant interaction between treatment and strain; E: Significant effect experimenter; n.s. = no significant difference. Significant *post hoc* comparisons indicated with a lower-case letter, where **(II-1)** the same lower-case letter above two single boxplots indicates a significant *post hoc* comparison between strains for one treatment condition only and **(II-2, III-1, III-2)** the same lower-case letter above combined boxplots indicates a significant *post hoc* comparison between strains regardless of treatment (*P*-value adjusted for multiple comparisons, S2 Table). The raw integrated z-scores (mean ± 95% CI) of each group depicted in this figure are presented in S7 Table.

https://doi.org/10.1371/journal.pone.0255521.g006

= 10.15, $P = 0.0032$, *very large* effect size, $\eta_P^2 = 0.241$, 95%CI [0.032, 0.450]; Fig 6III-1 and S7 Table) but when controlling for individual response type this treatment effect disappeared in the balanced pool ($F_{(1,32)} = 1.62$, $P = 0.2224$, *medium* effect size, $\eta_P^2 = 0.046$, 95%CI [0.000, 0.249]; Fig 6III-2 and S7 Table).

Similar to exploration, strains differed in locomotion in both the balanced ($F_{(2,32)} = 13.78$, $P = 0.0002$, *very large* effect size, $\eta_P^2 = 0.463$, 95%CI [0.176, 0.616]; Fig 6C-II) and the unbalanced pool ($F_{(2,32)} = 12.05$, $P = 0.0001$, *very large* effect size, $\eta_P^2 = 0.430$, 95%CI [0.144, 0.590]; Fig 6III-1). *Post hoc* comparisons (adjusted $\alpha = 0.02532$) showed that locomotion (rank transformed) was significantly higher in B6N than in C and 129S2 in both the unbalanced (C, $P = 0.0021$, *large* effect size, $d = -1.116$, 95%CI [-1.880, -0.354]; 129S2, $P < 0.0001$, *very large* effect size, $d = -1.713$, 95%CI [-2.520, -0.903]; Fig 6C-I and S6C Table) and balanced condition

(C, $P = 0.0002$, *large* effect size, $d = -1.320$, 95%CI [-2.090, -0.520]; 129S2, $P < 0.0001$, *very large* effect size, $d = -2.250$-, 95%CI [-3.290, -1.209], Fig 6III-2 and S6B Table).

In contrast to activity related behavior, avoidance behavior was not affected by treatment in both the unbalanced ($F_{(1,32)} = 1.33$, $P = 0.2573$, *medium* effect size, $\eta_P^2 = 0.040$, 95%CI [0.000, 0.023]; Fig 6I-1) and the balanced pool ($F_{(1,32)} = 2.18$, $P = 0.1499$, *medium* effect size, $\eta_P^2 = 0.064$, 95%CI [0.000, 0.277]; Fig 6I-2). Also, strains did not differ in avoidance behavior in either pool (unbalanced, $F_{(2,32)} = 1.31$, $P = 0.2830$, *medium* effect size, $\eta_P^2 = 0.076$, 95%CI [0.000, 0.272]; balanced, $F_{(2,32)} = 0.64$, $P = 0.5349$, *medium* effect size, $\eta_P^2 = 0.038$, 95%CI [0.000, 0.339]; Fig 6I-1 and 6II-2 and S7 Table).

The differences in results for exploration and locomotion show that the variation related to individual response types may augment observed treatment effects, as these effects disappeared when this variation was controlled for. Analysis of avoidance behavior in the unbalanced and the balanced pool however suggests that variation related to individual response type may also exert an opposite effect, in the sense that it may mask variation related to confounding factors. In the unbalanced pool, observed levels of avoidance behavior were not significantly different between Experimenters I and II ($F_{(1,32)} = 1.86$, $P = 0.1825$, *medium* effect size, $\eta_P^2 = 0.055$, 95% CI [0.012, 0.252]; Fig 6I-1 and S7 Table). Controlling for individual response type in the balanced condition however, resulted in significantly higher levels of observed avoidance behavior for Experimenter I than for Experimenter II ($F_{(1,32)} = 7.35$, $P = 0.0107$, *large* effect size, $\eta_P^2 = 0.180$, 95%CI [0.011, 0.399]; Fig 6I-2 and S7 Table). Experimenter effects were also found for exploratory activity, but now in both pools: Observed levels of exploration were higher in Experimenter II than in Experimenter I in the unbalanced pool ($F_{(1,32)} = 6.56$, $P = 0.0154$, *large* effect size, $\eta_P^2 = 0.168$, 95%CI [0.006, 0.384]; Fig 6II-1 and S7 Table) and the balanced pool ($F_{(1,32)} = 5.29$, $P = 0.0281$, *large* effect size, $\eta_P^2 = 0.135$, 95%CI [0.000, 0.356]; Fig 6II-1 and S7 Table).

4. Discussion

Matching experimental animals on their individual response type in control and test groups yielded different results than a comparable experimental pool in which these response types were not accounted for. These results demonstrate how including inter-individual variability in the composition of experimental groups may alter the observed pharmacological effects on behavioral performance. Also, by directly comparing a design in which inter-individual variability was accounted for (the balanced pool) versus a design in which this was not accounted for (the unbalanced pool) this study, to our knowledge, is the first to empirically demonstrate how this variability indeed may affect experimental outcomes.

As noted in the Introduction, the demonstration of a confounding effect of inter-individual variability in itself is not new, as several examples exist of how sub-populations within an experimental pool may mask the detection of overall group effects (i.e. increase the risk of a Type II error [14,19]. Such a masking effect was confirmed in the present experiment where matching individuals on their response type revealed a confounding experimenter effect for avoidance behavior that was not observed in the opposing design in which inter-individual variability was not accounted for. Also, inter-individual variability appeared to augment treatment effects for activity behavior (Fig 6). These results thus support the claim that a more active consideration of inter-individual variability in the design and outcomes of neurobehavioral preclinical research could contribute to the quality and reproducibility of experimental results [3].

In addition, this study demonstrated how data-driven analysis techniques, such as the clustering approach applied here, may facilitate an individual-based characterization of behavioral responses that encompasses the entire spectrum of variability in our data. As outlined in the

Introduction, the advantage of such a data-driven approach in animal models of behavioral dysfunction is that it more closely matches the conceptualization of human psychopathology on a continuous spectrum [26]. Thereby it provides a more refined alternative to other strategies that harness inter-individual variability by separating subpopulations based on a predefined criterion [14]. The present study also provided a methodological example of how inter-individual variability may be considered as a variable in the design of animal experiments. Interestingly, a highly similar approach was recently demonstrated by Rojas-Carvajal et al. [12], who first characterized unconditioned responses to novelty in individual Sprague-Dawley and Wistar outbred rats and subsequently used this information to balance experimental groups such that inter-individual variability in the read-out parameter of interest was equally distributed within and between experimental groups. This study, together with the one presented here, demonstrates how a priori characterization of individual experimental animals may enable researchers to actively control for any inter-individual variability in the design of their experiments.

At the same time however, we recognize that the procedure of a priori characterization may not be suitable or desirable for all phenotypes or contexts. In behavioral neurosciences for example, initial testing for the purpose of individual characterization may be undesirable because the animals need to be naïve to the test, although evidence suggests that naivety to the test can be ensured by allowing sufficient time between characterization and test [68,69]. Second, characterizing individuals and subsequently using this characterization in experimental design requires that the observed trait is consistent across time. With respect to anxiety, individuality in both anxiety-related and activity behaviors have indeed often been shown to be repeatable and consistent through time and context in isogenic mice [27–29,70–72]. Other behavioral traits however, such as grooming behavior, have been shown to be less consistent across time within the same individual [72]. In the present study we unfortunately did not test the temporal consistency of our individual response types. The identified clusters were stable however at the time of assessment (Fig 2) and the behavioral profiles of the response types largely overlapped with two previous studies [31,32], suggesting some consistency. This aspect requires further validation, as confirmation of temporal stability of our response types will substantiate our claim that the difference in results between a balanced and the unbalanced pool can be attributed to variation related to inter-individual differences in response.

Third, for some phenotypes, a priori characterization may simply not be possible because this phenotype is only expressed during a limited time window, such as the isolation calling response, or ultrasonic vocalization in rodent pups [73], or social play behavior in rats [74]. And, lastly, a priori characterization may present some serious challenges from a time and cost perspective. On that note, it should be emphasized that some of the more labor-intensive aspects of the present study (multiple trials for characterization of the animals, scoring multiple behavioral categories) were tied specifically to our phenotype of interest: temporal change of anxiety and activity related behavior. Other paradigms, assessing less complex phenotypes could suffice with less complex ethograms, a single trial, or even assessment in an automated home cage environment prior to testing.

A highly efficient alternative to a priori characterization is the use of cross-over designs [75,76]. In these designs, for example a balanced Latin Square design, variability between individuals is accounted for by contrasting all treatments within the same animal [76]. Each animal thereby serves as its own control, which reduces the sample size while maintaining the same statistical power [76]. With respect to the current study, it should be noted that this would indeed have been a more practical and time efficient strategy to account for inter-individual variability if evaluating the effectiveness of dexmedetomidine as an anxiolytic would have been the sole purpose of this paper. As described above however, our primary aim was not so much to study dexmedetomidine as an anxiolytic, but to evaluate whether balancing experimental

groups with respect to inter-individual variability would yield different results compared to a 'regular' drug study in which this variability is not actively accounted for.

Also, analogous to the strategy of a priori characterization, cross-over designs may not be suitable for all research contexts. A prerequisite for cross-over designs for example is that order effects are controlled for by balancing test order between subjects and that the applied treatment does not permanently alter the subjects [76]. In addition, similar to our approach, a sufficient wash-out period should be allowed between treatments to avoid carry-over effects [76]. Due to these requirements, cross-over designs are less suitable in studies that address the brain-behavior relationship in response to one or multiple compounds, as is common in research addressing the neurobiological underpinnings of behavioral dysfunction [77]. Similarly, these designs are less suitable in studies that evaluate chronic drug treatment, for example in the context of depression and anxiety [78,79]. All in all, both methods have their strengths in terms of controlling for inter-individual variability and what method is preferred highly depends on the research objective.

In addition to accounting for inter-individual variability in experimental design, one may also account for individual responses in statistical analysis of the results. In the present study, we used LMMs to analyze cluster differences and the effects of dexmedetomidine because these techniques have been proven especially effective in accounting for between- and within-individual variability [80]. The advantage of these models is that multiple characteristics of individual response (i.e. variability between and/or within individuals, differences in residual variation among individuals) can be accounted for in a single model [80,81]. LMMs also present a number of advantages over traditional analysis of variance (ANOVA/ANCOVA) when random effects are present, such as increased power, flexibility towards non-normally distributed data and the ability to handle missing values [82]. Accounting for such random effects in addition reduces the probability of false positives (Type I error rates) and false negatives (Type II error rate, [83]).

Next, in addition to these methodological and statistical considerations, this study confirmed previously identified inter-individual differences in the ability to habituate anxiety-related behavior in C, 129S2 and B6N [32]. The profiles were characterized by differential patterns of avoidance behavior and exploration: Mice in Cluster A combined an increase of avoidance behavior with low levels of exploration that remained stable over trials, while avoidance behavior decreased and exploration increased in Cluster B (Fig 3). This interplay between anxiety and exploration may be explained by the so-called approach-avoidance conflict that exposure to novel environments may induce in rodents, which entails the motivational conflict between the drive to explore a novel environment and the motivation to avoid potentially harmful stimuli [13,33,34]. Following this interplay, the decrease in avoidance behavior in Cluster B may be interpreted as successful habituation of anxiety responses, while the increase in avoidance in Cluster A may be considered as impaired habituation to the test.

In addition to contrasting patterns of avoidance behavior and exploration however, the clusters were characterized by significant differences in locomotor activity. The low levels of locomotion that decreased over trials in Cluster A contrasted with the pronounced increase in locomotion in Cluster B. Locomotion is not only associated with general activity levels, but differences in locomotion may also confound the interpretation of avoidance behavior as an indicator of anxiety, as the lack of exploration of an unprotected area may just as well be the result of reduced activity [33,34]. Whether the two individual response types reflect differential anxiety-phenotypes, or whether they merely reflect differential activity levels requires further study. A more elaborate discussion on this matter, along with suggestions for future research is provided in [32].

Furthermore, this study presented marked experimenter differences in behavioral scores for avoidance behavior and exploration, despite the fact that the experiment was carefully balanced with respect to experimenter. Behavioral phenotyping of anxiety-related behavior in rodents

has indeed been demonstrated to be sensitive to experimenter-related factors such as handling style and familiarity with the experimenter [84,85]. In this study, both experimenters were naïve to handling rodents and were trained by the same person to handle the mice. Furthermore, all animals were handled by both experimenters from arrival at the test facility onwards. These measures however unfortunately do not preclude the possibility that handling differences or other experimenter-related factors affected the outcome of this study as experimenters themselves can never be entirely subjected to standardization [86]. Another experimenter-induced factor that may affect the scoring of live behavior is observer variability [38,87,88]. Both inter- and intra-observer reliability were established at a good to excellent level prior to the start of the study, after a training phase in which both experimenters aligned coding by scoring video data from previously collected mHB-data. Observer reliability in itself may however again be affected by coding experience, rapidity of the behavior, energy level of the observer and so on [11,89]. These factors illustrate how complete control over experimenter-induced variability is difficult to accomplish. In fact, the experimenter has been termed one of the most uncontrollable background factors in experimental research, affecting experimental outcomes and reproducibility between studies in a similar manner as inter-individual variability [38].

Automated tracking may form a way to overcome this incontrollable nature and as such to increase the standardization of an experiment [88]. Fully automated scoring has unfortunately not yet been validated in the modified Hole Board however, and doing so was beyond the scope of the present study. It has been suggested however that experimenter effects should not greatly reduce the power to detect treatment effects provided the experiment was carefully balanced for the inclusion of multiple experimenters, and experimenter is included as a factor in the data analysis [87]–which was indeed the case in the present study. Also, systematic incorporation of multiple experimenters was recently suggested by Richter [38] as a means to account for potentially confounding experimenter-induced variation. This concept, termed systematic heterogenization, entails that one may improve the generalizability of results by systematically incorporating known sources of experimental variation (such as experimenter) in the design of a single experiment [8].

Finally, this study does not provide definitive conclusions about the potential of dexmedetomidine as an anxiolytic. Treatment with dexmedetomidine resulted in (a suggestion of) higher anxiety related behavior in treated animals, while exploration was significantly lower. This could be interpreted as anxiogenic according to the aforementioned interplay between anxiety-related behavior and exploration. We suspect however that the observed effects may rather have been associated with sedation, because locomotor activity was significantly lower in treated animals. Previous studies confirm a sedative effect of dexmedetomidine, as indicated by reduced locomotor activity [90,91]. Our choice for keeping the dose of dexmedetomidine constant across strains was motivated by our objective to keep factors other than individual response type the same between test groups. For a correct evaluation of the effect of dexmedetomidine however, inspecting strain-dependent dose-response behaviors would have probably been more appropriate as different mouse strains have demonstrated differences in a2-adrenergic receptor-binding [92,93]. This anxiogenic/sedative effect however did not interfere with the main objective of this paper.

Conclusions

This study empirically demonstrated that inter-individual variability may mask or augment experimental results. In addition, it provides an example approach of how this variability can be incorporated in experimental design, and how phenotypes that rely on the temporal nature of a response may be defined on an individual, multivariate level. As such it contributes to the

existing literature that explores new approaches and viewpoints in experimental design and analysis with the goal to improve the quality and reproducibility of experimental results.

Supporting information

S1 Table. Behavioral variables measured in mHB and used for composition of z-scores in this paper.
(DOCX)

S2 Table. Overview of Dunn-Sidak corrected values for α in *post hoc* comparisons.
(DOCX)

S3 Table. Mean integrated behavioral z-score and corresponding 95% confidence interval for each cluster (A/B) on each trial (1–5) for avoidance behavior, exploration and locomotion.
(DOCX)

S4 Table. Post hoc tests comparing either (I) the estimated marginal means between trials 1 and 5 (adjusted $\alpha = 0.016952$) for avoidance behavior, exploration and locomotion and (II) cluster differences on each trial for avoidance behavior, exploration and locomotion (adjusted $\alpha = 0.016952$). Significant comparisons are highlighted in bold.
(DOCX)

S5 Table. Raw integrated z-scores (mean ± 95% confidence interval) of groups (n = 8/group) that were compared in GLMM's to test the effects of treatment, strain, pool and experimenter on avoidance behavior, exploration and locomotor activity, using a 2 (treatment) x 3 (strain) x 2 (experimenter) x 2 (balanced/unbalanced pool) factorial design, including all interactions.
(DOCX)

S6 Table. Post hoc tests comparing (a) the estimated marginal means between strains (adjusted $\alpha = 0.025321$) for each behavioral dimension, on the total dataset, so balanced and unbalanced combined, section 2.2.1 (b) strain differences on each behavioral dimension (adjusted $\alpha = 0.025321$) for the balanced data only and (c) strain differences on avoidance behavior and locomotion (adjusted $\alpha = 0.025321$) or strain comparisons within treatment/within strain comparisons between treatments (adjusted $\alpha = 0.016952$) for exploration. Significant comparisons are highlighted in bold.
(DOCX)

S7 Table. Raw integrated z-scores (mean ± 95% confidence interval) of groups (*n* = 4/group) that were compared in GLMM's to test the effects of treatment, strain, pool and experimenter on avoidance behavior, exploration and locomotor activity, using a 2 (treatment) x 3 (strain) x 2 (experimenter) x 2 (balanced/unbalanced pool) factorial design, including all interactions.
(DOCX)

S1 Raw data.
(XLSX)

Acknowledgments

The authors would like to dedicate this article to the memory of prof. dr. Frauke Ohl. We remember Frauke as a beloved colleague and supervisor and we are most grateful for inspiring us to continue the work leading up to this manuscript. In addition, the authors are grateful to

Dr. J.R. Yates and an anonymous reviewer for their valuable comments on the manuscript, which has undoubtedly helped the authors to improve the article.

Author Contributions

Conceptualization: Marloes H. van der Goot, Saskia S. Arndt, Hein A. van Lith.

Data curation: Marieke Kooij, Suzanne Stolte, Annemarie Baars.

Formal analysis: Marloes H. van der Goot.

Investigation: Marloes H. van der Goot, Hein A. van Lith.

Methodology: Marloes H. van der Goot, Hein A. van Lith.

Project administration: Marloes H. van der Goot.

Supervision: Saskia S. Arndt, Hein A. van Lith.

Writing – original draft: Marloes H. van der Goot.

Writing – review & editing: Saskia S. Arndt, Hein A. van Lith.

References

1. Koolhaas JM, de Boer SF, Coppens CM, Buwalda B. Neuroendocrinology of coping styles: Towards understanding the biology of individual variation. Front. Neuroendocrinol. 2010; 31, 307–321. https://doi.org/10.1016/j.yfrne.2010.04.001 PMID: 20382177
2. Gärtner K. A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 yearlong effort to standardize laboratory animals? Int. J. Epidemiol. 2012; 41, 335–341. Reprint of Lab. Anim. 1990; 24, 71–77. https://doi.org/10.1093/ije/dyr219 PMID: 22266059
3. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in the light of biological variation. Nat. Rev. Neurosci. 2020; 21, 384–393. https://doi.org/10.1038/s41583-020-0313-3 PMID: 32488205
4. Lathe R. The individuality of mice. Genes Brain Behav. 2004; 3, 317–327. https://doi.org/10.1111/j.1601-183X.2004.00083.x PMID: 15544575
5. Einat H, Ezer I, Kara N, Belzung C. Individual responses of rodents in modelling of affective disorders and in their treatment: prospective review. Acta Neuropsychiatr. 2018; 30, 323–333. https://doi.org/10.1017/neu.2018.14 PMID: 29909818
6. Lewejohann L, Zipser B, Sachser N. „Personality"in laboratory mice used for biomedical research: A way of understanding variability? Dev. Psychobiol. 2001; 53 (6), 624–630.
7. Richter SH, Garner JP, Auer C, Kunert J, Wuerbel H. Systematic variation improves reproducibility of animal experiments. Nat. Meth. 2010; 7, 167–168. https://doi.org/10.1038/nmeth0310-167 PMID: 20195246
8. Richter SH. Systematic heterogenization for better reproducibility in animal experimentation. Lab. Anim. 2017; 46, 343–349. https://doi.org/10.1038/laban.1330 PMID: 29296016
9. Voelkl B, Würbel H. A reaction norm perspective on reproducibility. Preprint at http:/bioRxiv.org/content/10.1101/510941v3, 2020.
10. Kafkafi N, Agassi J, Chesler EJ. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci. Biobehav. Rev. 2018; 87, 218–232. https://doi.org/10.1016/j.neubiorev.2018.01.003 PMID: 29357292
11. Bello NM, Renter DG. Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. J. Dairy Sci. 2018; 101, 5679–5701. https://doi.org/10.3168/jds.2017-13978 PMID: 29729923
12. Rojas-Carvajal M, Quesada-Yamasaki D, Brenes JC. The cage test as an easy way to screen and evaluate spontaneous activity in preclinical neuroscience studies. Methodsx 2021; 8, 101271.
13. Armario A, Nadal R. Individual differences and the characterization of animal models of psychopathology: a strong challenge and a good opportunity. Front. Pharmacol. 2013; 4, 137. https://doi.org/10.3389/fphar.2013.00137 PMID: 24265618

14. Lonsdorf TB, Merz CJ. More than just noise: Inter-individual differences in fear acquisition, extinction and fear in humans–Biological, experiential, temperamental factors, and methodological pitfalls. Neurosci. Biobehav. Rev. 2017; 80, 703–728. https://doi.org/10.1016/j.neubiorev.2017.07.007 PMID: 28764976

15. Cohen H, Geva AB, Matar MA, Zohar J, Kaplan Z. Post-traumatic stress behavioural responses in inbred mouse strains: can genetic predisposition explain phenotypic variability? Int. J. Neuropsychoph. 2008; 11, 331–349.

16. Galatzer-Levy IR, Bonanno GA, Bush DEA, LeDoux JE. Heterogeneity in threat extinction learning: substantive and methodological considerations for identifying individual differences in response to stress. Front. Behav. Neurosci. 2013; 7, 55. https://doi.org/10.3389/fnbeh.2013.00055 PMID: 23754992

17. Pawlak CR, Ho Y, Schwarting RKW. Animal models of human psychopathology based on indivdiual differences in novelty-seeking and anxiety. Neurosci. Biobehav. Rev. 2008; 32, 1544–1568. https://doi.org/10.1016/j.neubiorev.2008.06.007 PMID: 18619487

18. Harro J. Inter-individual differences in neurobiology as vulnerability factors for affective disorders: Implications for psychopharmacology. Pharmacol. Ter. 2010; 125 (3), 402–422. https://doi.org/10.1016/j.pharmthera.2009.11.006 PMID: 20005252

19. Barbelivien A, Billy E, Lazarus C, Kelche C, Majchrzak M. Rats with different profiles of impulsive choice behavior exhibit differences in responses to caffeine and d-amphetamine and in medial prefrontal cortex 5-HT utilization. Behav. Brain. Res. 2008; 187 (2), 273–282. https://doi.org/10.1016/j.bbr.2007.09.020 PMID: 18029033

20. Festing MFW. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. ILAR Journal 2014; 55, 399–404. https://doi.org/10.1093/ilar/ilu036 PMID: 25541542

21. Festing MFW. Study Design. In: Martin-Kehl MI, Schubiger PA, editors. Animal Models for Human Cancer: Discovery and Development of Novel Therapeutics. Wiley-VCH, Weinheim; 2016, pp. 27–40. https://doi.org/10.1007/978-1-4939-3661-8_1 PMID: 27150081

22. Guiliano C, Peña-Liver Y, Goodlett CR, Cardinal RN, Robbins TW, Bullmore ET, et al. Evidence for a long-lastig compulsive alcohol seeking phenotype in rats. Neuropsychopharmacology 2018; 43, 728–738. https://doi.org/10.1038/npp.2017.105 PMID: 28553834

23. Giuliano C, Puaud M, Cardinal RN, Belin D, Everitt BJ. Individual differences in the engagement of habituation control over alcohol seeking predicts the development of compulsive alcohol seeking and drinking. Preprint at http:/bioRxiv.org/10.1111/adb13041, 2021.

24. Irwin JR, McClelland GH. Negative consequences of dichotomizing continuous predictor variables. J. Mark. Res. 2003; 40 (3), 366–371.

25. Stegman Y, Schiele MA, Schümann D, Lonsdorf TB, Zwanzger P, Romanos M, et al. Individual differences in human fear generalization–pattern identification and implications for anxiety disorders. Transl. Psychiatry 2019; 9, 307. https://doi.org/10.1038/s41398-019-0646-8 PMID: 31740663

26. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 2010; 167 (7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379 PMID: 20595427

27. Freund J, Brandmaier AM, Lewejohann L, Kirste I, Kritzler M, Krüger A, et al. Emergence of individuality in genetically identical mice. Science 2013; 340, 756–759. https://doi.org/10.1126/science.1235294 PMID: 23661762

28. Keshavarz M, Krebs-Wheaton R, Refki P, Savriama Y, Zhang Y, Guenther A, et al. Natural copy number variation differences of tandemly repeated small nucleolar RNAs in the Prader-Willi syndrome genomic region regulate individual behavioral responses in mammals. Preprint at http:/bioRxiv.org/content/10.1101/476010v2, 2020.

29. Kazavchinsky L, Dafna A, Einat H. Individual variability in female and male mice in a test-retest protocol of the forced swim test. J. Pharmacol. Toxicol. Methods 2019; 95, 12–15. https://doi.org/10.1016/j.vascn.2018.11.007 PMID: 30476619

30. Tuttle AH, Philip VM, Chesler EJ, Mogil JS. Comparing phenotypic variation between inbred and outbred mice. Nat. Methods 2018; 15 (12), 994–996. https://doi.org/10.1038/s41592-018-0224-7 PMID: 30504873

31. van der Goot MH, Boleij H, van den Broek J, Salomons AR, Arndt SS, van Lith HA. An individual based, multidimensional approach to identify emotional reactivity profiles in inbred mice. J. Neurosci. Meth. 2020; 343, 108810. https://doi.org/10.1016/j.jneumeth.2020.108810 PMID: 32574640

32. van der Goot MH, Keijsper M, Baars A, Drost L, Hendriks J, Kirchhoff S, et al. Inter-individual variability in habituation of anxiety-related responses within three mouse inbred strains. Phys. Behav. 2021; 113503 (Epub ahead of print). https://doi.org/10.1016/j.physbeh.2021.113503 PMID: 34153326

33. O'Leary TP, Gunn RK, Brown RE. What are we measuring when we test strain differences in anxiety in mice? Behav. Genet. 2013; 43 (1), 34–50. https://doi.org/10.1007/s10519-012-9572-8 PMID: 23288504

34. Ohl F. Testing for anxiety. Clin. Neurosci. Res. 2003; 3 (4–5), 233–238. https://doi.org/10.1046/j.1460-9568.2003.02436.x PMID: 12534976

35. Belzung C, Griebel G. Measuring normal and pathological anxiety-like behavior in mice: a review. Behav. Brain Res. 2001; 125 (1–2), 141–149. https://doi.org/10.1016/s0166-4328(01)00291-1 PMID: 11682105

36. Festing MFW. Experimental design and irreproducibility of pre-clinical research. Physiol. News 2020; 118, 14–15.

37. Festing MFW. The "completely randomized" and the "randomized block" are the only experimental designs suitable for widespread use in pre-clinical research. Sci. Rep. 2020; 10, 17577. https://doi.org/10.1038/s41598-020-74538-3 PMID: 33067494

38. Richter SH. Automated home cage testing as a tool to improve reproducibility of behavioral research? Front. Neurosci. 2020; 14, 383. https://doi.org/10.3389/fnins.2020.00383 PMID: 32390795

39. Salomons AR, Espitia Pinzon N, Boleij H, Kirchhoff S, Arndt SS, Nordquist RE, et al. Differential effects of diazepam and MPEP on habituation and neurobehavioral processes in inbred mice. Behav. Brain Funct. 2012; 8, 30. https://doi.org/10.1186/1744-9081-8-30 PMID: 22686184

40. Weerink MAS, Struys MMRF, Hannivoort LN, Barends CRM, Absalom AR, Colin P. Clinical pharmacokinetics and pharmacodynamics of dexmedetomidine. Clin. Pharmacokinet. 2017; 56, 893–913. https://doi.org/10.1007/s40262-017-0507-7 PMID: 28105598

41. Laarakker MC, van Raai JR, van Lith HA, Ohl F. The role of the alpha 2A-adrenoceptor in mouse stress-coping behaviour. Psychoneuroendicronology 2010; 35, 490–502. https://doi.org/10.1016/j.psyneuen.2009.08.014 PMID: 19766405

42. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. BMC Vet Res. 2020; 16, 242. https://doi.org/10.1186/s12917-020-02451-y PMID: 32660541

43. Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLos Biol. 2020; 18, e3000411. https://doi.org/10.1371/journal.pbio.3000411 PMID: 32663221

44. Meziane H, Quagazzal A-M, Aubert L, Wietrzych M, Krezel W. Estrous cycle effects on behavior of C57BL/6J and BALB/cByJ female mice: implications for phenotyping strategies. Genes Brain Behav. 2007; 6, 192–200. https://doi.org/10.1111/j.1601-183X.2006.00249.x PMID: 16827921

45. Arndt SS, Laarakker MC, van Lith HA, van der Staay FJ, Gieling E, Salomons AR, et al. Individual housing of mice–Impact on behavior and stress responses. Phys. Behav. 2009; 97 (3–4), 385–393.

46. Kappel S, Hawkins P, Mendl MT. To group or not to group? Good practice for housing male laboratory mice. Anim. 2017; 7 (12), 88.

47. Ohl F, Holsboer F, Landgraf R. The modified hole board as a differential screen for behavior in rodents. Behav. Res. Methods Instr. Comput. 2001; 33 (3), 392–397. https://doi.org/10.3758/bf03195393 PMID: 11591071

48. Labots M, van Lith HA, Ohl F, Arndt SS. The modified hole board–measuring behavior, cognition and social interaction in mice and rats. J. Vis. Exp. 2015; 98, e52529. https://doi.org/10.3791/52529 PMID: 25938188

49. Cicchetti DV. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsyc. 2001; 23, 695–700. https://doi.org/10.1076/jcen.23.5.695.1249 PMID: 11778646

50. Gertler R, Brown C, Mitchell D, Silvius E. Dexmedetomidine: a novel sedative-analgesic agent. Proc. (Bayl. Univ. Med. Cent.) 2001; 14, 13–21. https://doi.org/10.1080/08998280.2001.11927725 PMID: 16369581

51. Laarakker MC, Ohl F, van Lith HA. Chromosomal assignment of quantitaive trait loci influencing modified hole board behavior in laboratory mice using consomic strains, with special reference to anxiety-related behavior and mouse chromosome 19. Behav. Genet. 2008; 38, 159–184. https://doi.org/10.1007/s10519-007-9188-6 PMID: 18175213

52. Laarakker MC, van Lith HA, Ohl F. Behavioral characterization of A/J and C57BL/6J mice using a multidimensional test: association between blood plasma and brain magnesium-ion concentration with anxiety. Physiol. Behav. 2011; 102, 205–219. https://doi.org/10.1016/j.physbeh.2010.10.019 PMID: 21036185

53. Labots M, Laarakker MC, Schetters D, Arndt SS, van Lith HA. An improved procedure for integrated behavioral z-scoring illustrated with modified Hole Board behavior of male inbred laboratory mice. J. Neurosci. Methods 2018; 293, 375–388. https://doi.org/10.1016/j.jneumeth.2017.09.003 PMID: 28939008

54. R Core Team. R: A language environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria; 2020.

55. Pinheiro J, Bates D, Debroy S, Sarkar D, R Core Team. nlme: Linear and nonlinear mixed effects models. R package version 3.1.-147; 2020.

56. Genolini C, Alacoque X, Sentenac M, Arnaud C. Kml and kml3d: R-packages to cluster longitudinal data. J. Stat. Softw. 2015; 65, 1–34.

57. Sokal RR, Rohlf FJ. Biometry: The Principles and Practice of Statistics in Biological Research. 3rd ed. New York, NY: W.H. Freeman and Co. 1995.

58. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York, NY: Springer; 2005. https://doi.org/10.1007/s00405-005-0969-3 PMID: 16001247

59. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0; 2019.

60. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B. 2002; 63, 411–423.

61. Kryszczuk K, Hurley P. Estimation of the number of clusters using multiple clustering validity indices. In: El Gayar N, Kittler J, Roli F, editors. Multiple Classifier Systems. Lecture notes on Computer Science, vol 5997, New York, NY: Springer; 2010, pp 114–123.

62. Wahl S, Krug S, Then C, Kirchhofer A, Kastenmüller G, Brand T, et al. Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele. Metabolomics 2014; 10, 386–401.

63. Clatworthy J, Buick D, Hankins M, Weinman J, Home R. The use and reporting of cluster analysis in health psychology: a review. Br. J. Health. Psychol. 2005; 10, 329–358. https://doi.org/10.1348/135910705X25697 PMID: 16238852

64. Lenth R. Emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.7; 2020.

65. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. 1967; 62, 626–633.

66. Wahlsten D. Sample size. In: Wahlsten D, Mouse Behavioral Testing: How to Use Mice in Behavioral Neuroscience, London, UK: Academic Press, Elsevier Inc.; 2011, pp 75–105.

67. Festing MFW. The principles of experimental design and the determination of sample size when using animal models of traumatic brain injury. In: Srivastava A, Cox C, editors. Pre-Clinical and Clinical Methods in Brain Trauma Research, vol 139; New York, NY: Humana Press, 2018; pp. 201–225. https://doi.org/10.1177/0023677217738268 PMID: 29310487

68. Labots M, Zheng X, Moattari G, Ohl F, van Lith HA. Effects of light regime and substrain on behavioral profiles of male C57BL/6 mice in three tests of unconditioned anxiety. J. Neurogenet. 2016; 30, 306–315. https://doi.org/10.1080/01677063.2016.1249868 PMID: 27845603

69. Bouwknecht JA, Paylor R. Pitfalls in the interpretation of genetic and pharmacological effects on anxiety-like behaviour in rodents. Behav. Pharmacol. 2008; 19, 385–402. https://doi.org/10.1097/FBP.0b013e32830c3658 PMID: 18690100

70. Jakovcevski M, Schachner M, Morellini F. Individual variability in the stress response of C57BL/6J male mice correlates with trait anxiety. Genes Brain Behav. 2008; 7, 235–243. https://doi.org/10.1111/j.1601-183X.2007.00345.x PMID: 17680803

71. Montiglio P, Garant D, Thomas D, Réale D. Individual variation in temporal activity patterns in open-field tests. Anim. Behav. 2010; 80, 905–912.

72. Kalueff AV, Keisala T, Minasyan A, Kuuslahti M, Tuohimaa P. Temporal stability of novelty exploration in mice exposed to different open field tests. Behav. Proc. 2006; 72, 104–112. https://doi.org/10.1016/j.beproc.2005.12.011 PMID: 16442749

73. Hofer MA, Shair HN, Brunelli SA. Ultrasonic vocalizations in rat and mouse pups. Curr. Protoc. Neurosci. 2002; Chapter 8; Unit 8.14. https://doi.org/10.1002/0471142301.ns0814s17 PMID: 18428567

74. Vanderschuren LJMJ, Achterberg EJM, Trezza V. The neurobiology of social play and its rewarding value in rats. Neurosci. Biobehav. Rev. 2016; 70, 86–105. https://doi.org/10.1016/j.neubiorev.2016.07.025 PMID: 27587003

75. Martin P, Kraemer HC. Individual differences in behavior and their statistical consequences. Anim. Behav. 1987; 35 (5), 1366–1375.

76. Bate S, Clark R. Experimental Design. In: The Design and Statistical analysis of Animal Experiments, Cambridge UK: Cambridge University Press, pp. 30–121.

77. Van der Staay FJ. Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. Brain Res. Rev. 2006; 52, 131–159. https://doi.org/10.1016/j.brainresrev.2006.01.006 PMID: 16529820

78. Rodgers RJ, Cao BJ, Dalvi A, Holmes A. Animal models of anxiety: an ethological perspective. Braz. J. Med. Biol. Res. 1997; 30, 289–304. https://doi.org/10.1590/s0100-879x1997000300002 PMID: 9246227

79. Fuchs E, Flügge G. Experimental animal models for the simulation of depression and anxiety. Dialogues Clin. Neurosci. 2006; 8 (3), 323–333. https://doi.org/10.31887/DCNS.2006.8.3/efuchs PMID: 17117614

80. Bushby EV, Friel M, Goold C, Gray H, Smith L, Collins LM. Factors influencing individual variation in farm animal cognition and how to account for these statistically. Front. Vet. Sci. 2018; 5, 193. https://doi.org/10.3389/fvets.2018.00193 PMID: 30175105

81. Cleasby IR, Nakagawa S, Schielzeth H. Quantifying the predictability of behavior: statistical approaches for the study of between-individual variation and the within-individual variance. Methods Ecol. Evol. 2015; 6, 27–37.

82. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens HH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 2009; 24 (3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008 PMID: 19185386

83. Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, et al. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 2018; 6, e4794. https://doi.org/10.7717/peerj.4794 PMID: 29844961

84. Van Driel KS, Talling JC. Familiarity increases consistency in animal tests. Behav. Brain Res. 2005; 159, 243–245. https://doi.org/10.1016/j.bbr.2004.11.005 PMID: 15817187

85. Gouveia K, Hurst JL. Optimizing reliability of mouse performance in behavioral testing: the major role of non-aversive handling. Sci. Rep. 2017; 7, 44999. https://doi.org/10.1038/srep44999 PMID: 28322308

86. Lewejohann L, Reinhard C, Schrewe A, Brandewiede J, Haemisch A, Görtz N, et al. Environmental bias? Effects of housing conditions, laboratory environment and experimenter on behavioral tests. Genes Brain Behav. 2006; 5, 64–72. https://doi.org/10.1111/j.1601-183X.2005.00140.x PMID: 16436190

87. Bohlen M, Hayes ER, Bohlen B, Bailoo BD, Crabbe JC, Wahlsten D. Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. Behav. Brain Res. 2014; 217, 46–54. https://doi.org/10.1016/j.bbr.2014.06.017 PMID: 24933191

88. Spruijt BM, de Visser L. Advanced behavioral screening: automated home cage ethology. Drug Discov. Today, 2006; 3, 231–237.

89. Kaufman AB, Rosenthal R. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Anim. Behav. 2009; 78 (6), 1478–1491.

90. Salllinen J, Link RE, Haapalinna A, Viitamaa T, Kulatunga M, Sjöholm B, et al. Genetic alteration of alpha 2C-adrenoceptor expression in mice: influence on locomotor, hypothermic, and neurochemical effects of dexmedetomidine, a subtype-nonselective alpha 2-adrenoceptor agonist. Mol. Pharmacol. 1997; 51, 36–46. https://doi.org/10.1124/mol.51.1.36 PMID: 9016344

91. Votava M, Hess L, Slíva J, Kršiak M, Agová V. Dexmedetomidine selectively suppresses dominant behavior in aggressive and sociable mice. Eur. J. Pharmacol. 2005; 523, 79–85. https://doi.org/10.1016/j.ejphar.2005.08.022 PMID: 16226250

92. Fairbanks CA, Kitto KF, Nguyen HO, Stone LS, Wilcox GL. Clonidine and dexmedetomidine produce antinociceptive synergy in mouse spinal cord. Anesthesiology 2009; 110, 648–647. https://doi.org/10.1097/ALN.0b013e3181974f7a PMID: 19212263

93. Wilkinson M, Manchester EL. Strain differences in brain alpha2 and beta-adrenergic receptor binding in dystrophic mice. Brain Res. Bull. 1983; 11, 743–745. https://doi.org/10.1016/0361-9230(83)90018-7 PMID: 6318921

Table S1. Behavioral variables measured in mHB and used for composition of z-scores in this paper.

Motivational system/Behavioral dimension	*Behavioral variable*	*Directionality z-score*[1]
Anxiety related behavior		
- *Avoidance behavior*	Total number of board entries	-z
	Latency until first board entry	z
	Percentage of time spent on the board	-z
Activity		
- *Exploration*	Total number of rearings in the box	z
	Latency until first rearing in the box	-z
	Total number of rearings on the board	z
	Latency until first rearing on the board	-z
	Total number of hole explorations	z
	Latency until first hole exploration	-z
	Total number of hole visits	z
	Latency until first hole visit	-z
- *Locomotion*	Total number of line crossings	z
	Latency until first line crossing	-z

[1] Directionality of z-score: z-scores were adjusted as such that increase of value reflects increase in corresponding behavioral dimension: [Z]=regular z-score; [-Z]=adjusted z-score.

Table S2. Overview of Dunn-Sidak corrected values for α in *post hoc* comparisons.

Results section	Analysis type	GLMM effect	Post hoc comparisons/contrasts	γ	Adjusted α
2.1. Cluster analyses	LMM	Cluster (A/B) x Trial (T)	A-T1 vs A-T5; B-T1 vs B-T5; A-T1 vs B-T1; A-T5 vs B-T5	2	0.025321
	LMM		A-T2 vs B-T2; A-T3 vs B-T3; A-T4 vs B-T4	1	0.05
2.2.1	GLM	Strain	C vs B6N; C vs 129S2; B6N vs 129S2	2	0.025321
2.2.2	GLM	Strain	C vs B6N; C vs 129S2; B6N vs 129S2	2	0.025321
	GLM	Strain x Treatment (1 = dex/0 = saline)	C-1 vs C-0; B6N-1 vs B6N-0; 129S2-1 vs 129S2-0; C-1 vs B6N-1; C-1 vs 129S2-1; B6N-1 vs 129S2-1; C-0 vs B6N-0; C-0 vs 129S2-0; B6N-0 vs 129S2-0	3	0.01692

Table S3. Mean integrated behavioral z-score and corresponding 95% confidence interval for each cluster (A/B) on each trial (1-5) for avoidance behavior, exploration and locomotion.

		Cluster A			Cluster B		
Dimension	*trial*	*mean*	*ci_lower*	*ci_upper*	*mean*	*ci_lower*	*ci_upper*
Avoidance behavior	1	-0.24308	-0.41723	-0.06892	0.585057	0.418256	0.751857
	2	-0.46376	-0.60747	-0.32005	0.114297	-0.05352	0.282115
	3	0.035222	-0.11867	0.189111	-0.13797	-0.30511	0.029171
	4	0.309293	0.163164	0.455422	-0.325	-0.48178	-0.16823
	5	0.483182	0.34714	0.619224	-0.40019	-0.55286	-0.24751
Exploration	1	-0.1861	-0.25953	-0.11267	-0.37601	-0.4554	-0.29662
	2	0.066687	-0.04843	0.1818	-0.06227	-0.1757	0.051154
	3	-0.02464	-0.14377	0.094501	0.215273	0.083324	0.347222
	4	-0.18049	-0.30064	-0.06034	0.452915	0.302902	0.602928
	5	-0.27097	-0.39648	-0.14545	0.577163	0.43668	0.717645
Locomotion	1	-0.06232	-0.21351	0.088868	-0.20182	-0.40712	0.003474
	2	-0.09902	-0.26566	0.067618	0.275214	0.092459	0.457969
	3	-0.21974	-0.39384	-0.04563	0.438686	0.329698	0.547674
	4	-0.27481	-0.44041	-0.10921	0.456813	0.343232	0.570394
	5	-0.38321	-0.56038	-0.20605	0.439368	0.302839	0.575897

Table S4. *Post hoc* tests comparing either (I) the estimated marginal means between trials 1 and 5 (adjusted α = 0.016952) for avoidance behavior, exploration and locomotion and (II) cluster differences on each trial for avoidance behavior, exploration and locomotion (adjusted α = 0.016952). Significant comparisons are highlighted in bold.

(I) Dimension		**Estimate ± SEM**	$t_{(df)}$	***P***	**Cohens *d* [95% CI]**
Avoidance					
Trial 1 vs 5					
	A	-0.726 ± 0.96	$-7.593_{(708)}$	**< 0.0001**	-1.015 [-1.283, -0.747]
	B	0.985 ± 0.96	$10.288_{(708)}$	**< 0.0001**	1.377 [1.105, 1.650]
Exploration					
Trial 1 vs 5					
	A	0.085 ± 0.05	$1.583_{(708)}$	0.1138	0.319 [-0.077, 0.716]
	B	-0.953 ± 0.06	$-15.274_{(708)}$	**< 0.0001**	-3.588 [-4.085, -3.090]
Locomotion					
(rank transformed)					
Trial 1 vs 5					
	A	98.06 ± 23.1	$4.250_{(708)}$	**< 0.0001**	0.498 [0.267, 0.730]
	B	-252.68 ± 32.6	$-7.740_{(708)}$	**< 0.0001**	-1.285 [-1.618, -0.952]
(II) Dimension		**Estimate ± SEM**	$t_{(df)}$	***P***	**Cohen's *d* [95% CI]**
Avoidance					
A vs B	Trial 1	-0.828 ± 0.111	$-7.430_{(177)}$	**< 0.0001**	-1.158 [-1.488, -0.827]
	Trial 2	-0.578 ± 0.111	$-5.186_{(177)}$	**< 0.0001**	-0.808 [-1.127, -0.489]
	Trial 3	0.173 ± 0.111	$1.554_{(177)}$	0.1220	0.242 [-0.066, 0.551]
	Trial 4	0.634 ± 0.111	$5.691_{(177)}$	**< 0.0001**	0.887 [0.565, 1.208]
	Trial 5	0.883 ± 0.111	$7.926_{(177)}$	**< 0.0001**	1.235 [0.901, 1.569]
Exploration					
A vs B	Trial 1	0.190 ± 0.057	$3.318_{(177)}$	**0.0011****	0.715 [0.283, 1.147]

	Trial 2	0.129 ± 0.078	$1.658_{(177)}$	0.0992	0.485 [-0.095, 1.066]
	Trial 3	-0.240 ± 0.083	$-2.900_{(177)}$	**0.0042****	-0.903 [-1.525, -0.281]
	Trial 4	-0.633 ± 0.095	$-6.658_{(177)}$	**< 0.0001**	-2.384 [-3.134, -1.635]
	Trial 5	-0.848 ± 0.100	$-8.452_{(177)}$	**< 0.0001**	-3.192 [-4.009, -2.375]
Locomotion (rank transformed)					
A vs B	Trial 1	43.7 ± 46.0	$0.949_{(177)}$	0.3438	0.222 [-0.240, 0.685]
	Trial 2	-152.1 ± 35.7	$-4.260_{(177)}$	**< 0.0001**	-0.774 [-1.140, -0.406]
	Trial 3	-227.4 ± 33.6	$-6.774_{(177)}$	**< 0.0001**	-1.157 [-1.510, -0.798]
	Trial 4	-272.1 ± 31.8	$-8.551_{(177)}$	**< 0.0001**	-1.384 [-1.730, -1.033]
	Trial 5	-307.1 ± 31.2	$-9.851_{(177)}$	**< 0.0001**	-1.562 [-1.910, -1.208]

Table S5. Raw integrated z-scores (mean ± 95% confidence interval) of groups (n=8/group) that were compared in GLMM's to test the effects of treatment, strain, pool and experimenter on avoidance behavior, exploration and locomotor activity, using a 2 (treatment) x 3 (strain) x 2 (experimenter) x 2 (balanced/unbalanced pool) factorial design, including all interactions.

Dimension	Main effect/condition	n	mean	ci_lower	ci_upper
Avoidance behavior	treatment - saline	48	-0.16072	-0.3636	-0.52433
	treatment - dex	48	0.160721	-0.11315	0.047571
	strain - 129S2	32	-0.2489	-0.5847	-0.8336
	strain - C	32	-0.02607	-0.33485	-0.36093
	strain - B6N	32	0.274979	0.039869	0.510089
	pool - unbalanced	48	0.188684	-0.06204	0.126641
	pool - balanced	48	-0.18868	-0.41588	-0.60457
	experimenter - I	48	0.228397	-0.0024	0.225996
	experimenter - II	48	-0.2284	-0.46998	-0.69838
Exploration	treatment - saline	48	0.145162	-0.0353	0.109859
	treatment - dex	48	-0.14516	-0.29899	-0.44415
	strain - 129S2	32	-0.29773	-0.39063	-0.68837
	strain - C	32	-0.10129	-0.28281	-0.3841
	strain - B6N	32	0.39902	0.146633	0.545652
	pool - unbalanced	48	-0.10192	-0.25331	-0.35523
	pool - balanced	48	0.101924	-0.08557	0.016359
	experimenter - I	48	-0.15937	-0.30941	-0.46877
	experimenter - II	48	0.159367	-0.02222	0.137148
Locomotion	treatment - saline	48	0.216326	0.032184	0.24851
	treatment - dex	48	-0.21633	-0.48164	-0.69796
	strain - 129S2	32	-0.46171	-0.76736	-1.22907
	strain - C	32	-0.057	-0.27732	-0.33432
	strain - B6N	32	0.518705	0.283088	0.801793
	pool - unbalanced	48	0.021259	-0.2095	-0.18825
	pool - balanced	48	-0.02126	-0.26421	-0.28547
	experimenter - I	48	-0.13789	-0.42482	-0.5627
	experimenter - II	48	0.137889	-0.02568	0.112207

Table S6. *Post hoc* tests comparing (a) the estimated marginal means between strains (adjusted α = 0.025321) for each behavioral dimension, on the total dataset, so balanced and unbalanced combined, section 2.2.1 (b) strain differences on each behavioral dimension(adjusted α = 0.025321) for the balanced data only and (c) strain differences on avoidance behavior and locomotion (adjusted α = 0.025321) or strain comparisons within treatment/within strain comparisons between treatments (adjusted α = 0.016952) for exploration. Significant comparisons are highlighted in bold.

(a) Balanced and unbalanced pool combined: post hoc comparisons					
Dimension		Estimate ± SEM	z	P	Cohens d [95% CI]
Avoidance					
Strain effect	129S2 vs C	-0.192 ± 0.205	-0.940	0.3471	-0.239 [-0.738, 0.260]
	129S2 vs B6N	-0.465 ± 0.213	-2.184	0.0290	-0.577 [-1.104, -0.050]
	C vs B6N	-0.273 ± 0.204	-1.335	0.1819	-0.338 [-0.838, 0.162]
Exploration					
Strain effect	129S2 vs C	-0.220 ± 0.140	-1.571	0.1161	-0.399 [-0.901, 0.103]
	129S2 vs B6N	-0.730 ± 0.145	-5.024	**< 0.0001**	-1.327 [-1.891, -0.763]
	C vs B6N	-0.511 ± 0.139	-3.663	**0.0002**	-0.928 [-1.499, -0.408]
Locomotion (rank transformed)					
Strain effect	129S2 vs C	-15.8 ± 5.28	-2.994	**0.0028**	-0.760 [-1.270, -0.246]
	129S2 vs B6N	-41.6 ± 5.50	-7.563	**< 0.0001**	-2.000 [-2.620, -1.381]
	C vs B6N	-25.8 ± 5.27	-4.886	**< 0.0001**	-1.240 [-1.780, -0.700]
(b) Balanced pool only: post hoc comparisons					
Dimension		Estimate ± SEM	z	P	Cohens d [95% CI]
Avoidance					
Strain effect	129S2 vs C	-0.235 ± 0.316	-0.744	0.4568	-0.317 [-1.155, 0.521]
	129S2 vs B6N	-0.377 ± 0.333	-1.130	0.2587	-0.507 [-1.396, 0.382]
	C vs B6N	-0.141 ± 0.264	-0.535	0.5927	-0.190 [-0.889, 0.509]
Exploration					
Strain effect	129S2 vs C	-0.352 ± 0.225	-1.563	0.1181	-0.655 [-1.520, 0.185]
	129S2 vs B6N	-0.909 ± 0.237	-3.828	**0.0001**	-1.719 [-2.690, -0.743]
	C vs B6N	-0.557 ± 0.188	-2.961	**0.0031**	-1.053 [-1.800, -0.310]
Locomotion (rank transformed)					
Strain effect	129S2 vs C	-9.86 ± 4.51	-2.183	0.0290	-0.93 [-1.79, -0.06]
	129S2 vs B6N	-23.83 ± 4.76	-5.006	**< 0.0001**	-2.25 [-3.29, -1.209]
	C vs B6N	-13.98 ± 3.77	-3.702	**0.0002**	-1.32 [-2.09, -0.550]
(c) Unbalanced pool only: post hoc comparisons					
Dimension		Estimate ± SEM	z	P	Cohens d [95% CI]
Avoidance					
Strain effect					
	129S2 vs C	-0.088 ± 0.331	-0.267	0.7893	-0.096 [-0.805, 0.612]
	129S2 vs B6N	-0.495 ± 0.324	-1.528	0.1266	-0.540 [-1.246, 0.165]
	C vs B6N	-0.407 ± 0.333	-1.223	0.2215	-0.444 [-1.163, 0.276]
Exploration					
Strain x Treatment					

interaction					
Treatment (T)	129S2 vs C	-0.012 ± 0.219	0.056	0.9556	-0.028 [-0.961, 1.018]
	129S2 vs B6N	-0.153 ± 0.217	-0.708	0.4793	-0.354 [-1.338, 0.630]
	C vs B6N	-0.166 ± 0.220	-0.753	0.4514	-0.382 [-1.381, 0.617]
Control (C)	129S2 vs C	-0.273 ± 0.219	-1.245	0.2133	-0.630 [-1.634, 0.374]
	129S2 vs B6N	-1.097 ± 0.217	-5.062	**< 0.0001**	-2.531 [-3.691, -1.371]
	C vs B6N	-0.824 ± 0.219	-3.757	**0.0002**	-1.901 [-2.997, -0.805]
C versus T	129S2	-0.031 ± 0.217	-0.144	0.8852	-0.072 [-1.052, 0.908]
	C	0.254 ± 0.217	1.171	0.2415	0.586 [-0.405, 1.576]
	B6N	0.913 ± 0.217	4.207	**< 0.0001**	2.105 [0.997, 3.213]
Locomotion (rank transformed)					
Strain effect	129S2 vs C	-6.36 ± 3.85	-1.653	0.0983	-0.597 [-1.32, 0.126]
	129S2 vs B6N	-18.26 ± 3.77	-4.845	**< 0.0001**	-1.713 [-2.52, -0.903]
	C vs B6N	-11.89 ± 3.87	-3.076	**0.0021**	-1.116 [-1.88, -0.354]

Table S7. Raw integrated z-scores (mean ± 95% confidence interval) of groups (*n* = 4/group) that were compared in GLMM's to test the effects of treatment, strain, pool and experimenter on avoidance behavior, exploration and locomotor activity, using a 2 (treatment) x 3 (strain) x 2 (experimenter) x 2 (balanced/unbalanced pool) factorial design, including all interactions.

			Design: balanced			Design: *not* balanced		
Dimension	Experimenter	Strain/treatment	mean	ci_lower	ci_upper	mean	ci_lower	ci_upper
Avoidance behavior	Exp I	129S2 -saline	0.023019	-1.61607	1.662108	0.200246	-1.38651	1.787001
		129S2-dex	0.129693	-1.43332	1.692703	0.016335	-1.94767	1.980338
		C-saline	0.101698	-1.47017	1.673563	-0.03734	-1.26342	1.188748
		C-dex	0.441862	-0.82784	1.711569	0.193076	-1.60137	1.987524
		B6N-saline	0.429354	-0.13032	0.989028	0.09654	-0.59096	0.784039
		B6N-dex	0.625128	-0.28483	1.535088	0.571633	-0.73361	1.876876
	Exp II	129S2 -saline	-0.69476	-1.12455	-0.26498	-0.37338	-1.97587	1.229113
		129S2-dex	-0.73457	-1.73569	0.266541	-0.59157	-2.28593	1.102792
		C-saline	-0.55749	-1.06447	-0.05052	-0.5685	-1.24379	0.10679
		C-dex	0.316467	-1.34841	1.981347	-0.08268	-2.19269	2.027338
		B6N-saline	-0.27818	-0.78015	0.223796	-0.24184	-0.77131	0.287622
		B6N-dex	0.197786	-1.37734	1.772915	0.817472	0.487122	1.147822
Exploration	Exp I	129S2 -saline	-0.39387	-0.87417	0.086425	-0.4229	-0.89838	0.052583
		129S2-dex	-0.50552	-0.85873	-0.1523	-0.43842	-0.92738	0.050539
		C-saline	-0.25477	-0.88672	0.377188	-0.222	-0.76146	0.317457
		C-dex	-0.33452	-0.79876	0.129717	-0.32847	-0.7302	0.07327
		B6N-saline	0.187644	-0.33855	0.713839	0.767557	-0.60851	2.143629
		B6N-dex	0.242245	-0.85694	1.341427	-0.23107	-0.678	0.21587
	Exp II	129S2 -saline	-0.16467	-0.31148	-0.01785	-0.13366	-0.59896	0.33163
		129S2-dex	-0.24404	-0.48126	-0.00681	-0.05553	-0.21117	0.100119
		C-saline	0.352888	-0.74396	1.449732	0.287712	-0.26056	0.835983
		C-dex	-0.1803	-1.0436	0.682995	-0.12176	-1.29527	1.051738
		B6N-saline	0.8984	-0.08546	1.882258	0.870524	0.261631	1.479417
		B6N-dex	0.396512	-1.26627	2.05929	0.028014	-0.48706	0.543084

Locomotion	Exp I	129S2 -saline	-0.30246	-0.84295	0.238036	-0.03936	-0.55794	0.479215
		129S2-dex	-1.14415	-3.43442	1.146118	-1.3907	-3.63961	0.858205
		C-saline	-0.14697	-1.39312	1.099179	0.038583	-0.8603	0.937462
		C-dex	-0.21732	-2.33342	1.898771	0.025043	-0.884	0.934084
		B6N-saline	0.851491	0.214075	1.488907	0.57257	0.083174	1.061966
		B6N-dex	0.473804	-0.11642	1.064024	-0.39373	-2.17827	1.390809
	Exp II	129S2 -saline	-0.35352	-1.28484	0.57779	-0.11184	-0.58074	0.357052
		129S2-dex	-0.25024	-0.53849	0.038012	-0.06912	-0.55721	0.418971
		C-saline	-0.01579	-0.37472	0.343136	0.194522	-0.21018	0.599226
		C-dex	0.012291	-0.74927	0.773856	-0.34784	-0.83812	0.142431
		B6N-saline	0.875258	0.393329	1.357187	1.045723	-0.19156	2.283011
		B6N-dex	0.217619	-0.45245	0.887684	0.476164	0.132941	0.819387