Physiology & Behavior 239 (2021) 113503

Contents lists available at ScienceDirect

Physiology & Behavior

journal homepage: www.elsevier.com/locate/physbeh

Inter-individual variability in habituation of anxiety-related responses within three mouse inbred strains

Marloes H. van der Goot [a,b,*], Melissa Keijsper [a], Annemarie Baars [a], Lisa Drost [a], Judith Hendriks [a], Susanne Kirchhoff [a], José G. Lozeman-van t Klooster [a], Hein A. van Lith [a,b], Saskia S. Arndt [a]

[a] Section Animals in Science and Society, Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands
[b] Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Inter-individual variability
Inbred mice
Anxiety
Habituation
Cluster analysis

ABSTRACT

Inter-individual variability in behavioral and physiological response has become a well-established phenomenon in animal models of anxiety and other disorders. Such variability is even demonstrated within mouse inbred strains. A recent study showed that adaptive and non-adaptive anxiety phenotypes (measured as habituation and/or sensitization of anxiety responses) may differ within cohorts of 129 mice. This variability was expressed across both anxiety- and activity-related behavioral dimensions. These findings were based however on re-analysis of previously published data. The present study therefore aimed to empirically validate these findings in 129 mice. In addition, we assessed such inter-individuality in two other strains: BALB/c and C57BL/6.

Males of three mouse inbred strains (BALB/c, C57BL/6 and 129S2) were behaviorally characterized through repeated exposure to a mild aversive stimulus (modified Hole Board, 4 consecutive trials). Behavioral observations were supplemented with assessment of circulating corticosterone levels.

Clustering the individual response trajectories of behavioral and endocrine responses yielded two multidimensional response types of different adaptive value. Interestingly, these response types were displayed by individuals of all three strains. The response types differed significantly on anxiety and activity related behavioral dimensions but not on corticosterone concentrations.

This study empirically confirms that adaptive capacities may differ within 129 cohorts. In addition, it extends this inter-individual variability in behavioral profiles to BALB/c and C57BL/6. Whether these two sub-types constitute differential anxiety phenotypes may differ per strain and requires further study.

Abbreviations

mHB	modified Hole Board
pCORT	blood plasma corticosterone concentrations (nmol/L)
CVI	Clustering Validity Index; GLMM: Generalized linear mixed models

1. Introduction

Inter-individual variability in emotional reactivity to environmental challenges is a well-established phenomenon in animal models of stress, anxiety, depression, and post-traumatic stress disorder (e.g. [1][2][3][4][5][6]). This type of variability has repeatedly been associated with complex interactions between genetic and environmental factors, which are partly modulated by epigenetic processes [7][8].

Inter-individual variability has become of increasing interest in animal models that study the underlying mechanisms and/or treatment of psychiatric diseases [3][6]. In humans, the susceptibility to develop psychopathologies and the response to treatment is known to vary greatly between patients [3]. Similar circumstances may trigger development of affective disorders in some, while other individuals are unaffected [9].

Incorporating this variation in animal models may therefore not only make these models more representative [6], but could also improve our understanding of the underlying mechanisms that are involved in this differential susceptibility [3][9].

* Corresponding author. Section Animals in Science and Society, Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Building Nieuw Gildestein, Yalelaan 2, 3584 CM Utrecht, the Netherlands. Tel: 0031-648251689.
E-mail address: m.h.vandergoot@uu.nl (M.H. van der Goot).

https://doi.org/10.1016/j.physbeh.2021.113503
Received 15 January 2021; Received in revised form 26 March 2021; Accepted 16 June 2021
Available online 18 June 2021

It has been suggested that a starting point for mapping such differential susceptibility could be a more in depth, individual based, characterization of behavior of a particular animal model [3]. This could then be followed by the identification of biological markers that may explain the differences between these sub-groups [3][9].

Behavioral habituation and sensitization are two contrasting forms of learning, and are defined as either the decremental (habituation) or incremental (sensitization) change in behavioral response after repeated exposure to environmental stimuli (provided these stimuli are not accompanied by biologically significant consequences) [10].

In rodents, exposure to novelty induces a biologically adaptive anxiety response that enables individuals to respond appropriately to potential threat [11]. In an adaptive phenotype, repeated exposure to such stimuli results in habituation (i.e. the waning) of anxiety-related behavior, enabling individuals to adapt to environmental challenges [11][14]. Several studies suggest that the opposite of a habituation response (i.e. a sensitization of anxiety behavior) may reflect a non-adaptive anxiety phenotype, that can be used as an indicator of pathological anxiety in rodent models [11][13][14][15][16].

In these studies, two strains that differed in innate emotionality [17] were repeatedly exposed to a mild aversive stimulus (the modified Hole Board). BALB/cJ (known as a neophobic mouse strain [18]) displayed initial high levels of anxiety that decreased with repeated exposure to the test. At the same time, exploration and locomotion increased, which, taken together, reflected successful habituation. In contrast, 129 mice consistently showed low initial levels of anxiety-related behavior that increased with repeated exposure, suggesting a sensitizing anxiety response [11][13][14][15][16]. This non-adaptive behavioral profile was further characterized by a lower expression of the immediate early gene c-Fos (a marker for neural activity) in the prelimbic cortex and lateral septum, brain areas involved in the integration of emotional and cognitive processes, compared to rapidly habituating BALB/c.

Re-inspection of the behavioral data from these studies however demonstrated that responses may differ between individuals within these strains [19]. Using a multivariate cluster analysis on the combined data from these studies, van der Goot et al. [19] identified two homogenous subgroups of mice that followed the same response across trials: a habituation and a sensitization cluster. These clusters were found to be multidimensional, with individual mice consistently grouping together across dimensions indicative of anxiety related behavior, but also activity behavior [19].The profiles of these subtypes mirrored the BALB/c specific habituation response, and the sensitization response that was characteristic for 129 mice. These individual based analyses however also revealed that a sub-population of 129 mice displayed a successful habituation response, which was overlooked when only comparing average strain responses.

The identified behavioral profiles were based on retrospect analyses on a dataset that consisted of multiple experiments. These studies were conducted over a time span of 4 years and varied naturally in experimenter, test location, time of year etcetera – all factors that are known to affect variability between experiments [20][21].

Therefore, the question remained to what extent the observed inter-individual variability, and its expression within and across strains was representative of variation in (sub-)types of response BALB/c and 129 mice in general, or whether the identified clusters were part result of the mere variation that was inherent to analyzing a dataset consisting of multiple experiments.

The goal of the present experiment therefore was to empirically validate the previous findings, and assess inter-individuality in adaptive capacities in a controlled experiment. Three mouse inbred strains were behaviorally characterized by repeated exposure to the modified Hole Board. Two strains - BALB/c and 129S2 - were also observed in the previous studies [13][14][15][16]. In addition, we assessed inter-individual variability in habituation and sensitization responses in an additional strain: C57BL/6. According to the data presented by The Jackson Laboratory these three inbred lines are the most frequently used mouse strains in biomedical research. Phenotypic characteristics of these strains have been reviewed elsewhere [22]. C57BL/6 mice are typically classified as non-anxious and highly active [23][24][25][26]. When repeatedly exposed to the modified Hole Board, these mice are characterized as highly active, displaying low levels of anxiety-related behavior and showing no further habituation to the test [27].

Mice were individually characterized on the same five behavioral dimensions that comprised the previously identified clusters [19]: avoidance behavior, risk assessment, arousal, exploration and locomotion.

Alongside these behavioral responses we assessed whether individual variation on a behavioral level would also be reflected in corticosterone concentrations. In general, circulating corticosterone levels have been found to correlate with the intensity of anxiogenic situations in mice [28]. Glucocorticoid responses however, may also vary considerable in response to stressors in rodents [29][30][31]. High trait anxiety for example, correlated positively with plasma corticosterone concentrations in sub populations of C57BL/6 mice [32][33][34]. Behavioral observations were therefore supplemented with the assessment of plasma glucocorticoid concentrations.

2. Materials and methods

2.1. Ethical statement

The experimental protocol was approved by the Central Animal Experiments Committee, Den Haag, the Netherlands (CCD approval numbers: AVD1080020172264 and AVD1080020172264–1). The decision for approval was based on 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 [35][36].

2.2. Animals and housing

This study tested naïve males of three mouse inbred strains: BALB/cAnNCrl (hereafter C, see http://www.informatics.jax.org/mgihome/nomen/strains.shtml#labcodes, $n = 40$, albino), C

57BL/6NCrl (B6N, $n = 40$, black) and 129S2/SvPasCrl (129S2, $n = 38$, agouti). An additional two 129S2 mice died due to health reasons unrelated to the study and were not tested. The sample size was determined beforehand and based on recommendations for cluster analysis by Dolnicar et al. [37].

This study assessed inter-individuality in male mice only. It has reliably been established that incorporating both sexes when analyzing between group factor (strain, sex etc.) and/or treatment effects in factorial designs does not necessitate a duplication of sample size [39][40][41].

To our knowledge however, it is unclear whether the same mechanism applies to unsupervised clustering techniques, such as the one used in this study. Unsupervised clustering approaches do not have any a priori assumptions regarding the distribution and variance of the data [40]. This unsupervised nature makes these analyses more sensitive to anomalies in the data than classical statistical approaches, and it has been established that the detection of meaningful clusters increases with sample size [37]. To ensure a maximal sample size, while maintaining a manageable technical load due to repeated testing in multiple inbred strains, it was therefore decided to first explore this variability in males, with the intention to extend potential findings to females in a follow up study.

Animals were bred by and purchased from Charles River Germany (Sulzfeld, Germany) and arrived at the research facility in four batches of $n = 10$ per strain. All mice were 6–8 weeks old upon arrival (mean

body weight ± SEM and range of strains per batch are in supplementary Table S1). All 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 in Macrolon Type II L cages (size: 365 × 207 × 140 mm, floor area 530cm^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 PVC-shelter (Plexx BV, Elst, the Netherlands) as enrichment. Our previous research demonstrated that stress-levels in individually housed male mice did not differ significantly from socially housed male mice [38]. 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 kept similar between testing rooms. Relative humidity (mean percentage ± SD) was controlled (room A: 61.8% ± 3.93, range 48.3% – 81.0%; room B: 62.4% ± 4.16, range 50.5% – 81.1%) with a ventilation rate of 15–20 changes/hour and an average room temperature (mean *C* ± SD) of 22.1 *C*± 0.33 (range 20.4 – 23.6) and 22.2 *C*± 0.33 (range 20.3 – 23.7) for rooms A and B respectively. Both animal rooms also housed female C57BL/6NCrl mice throughout the entire duration of the study.

The mice were handled three times a week during the habituation period by the same experimenters that conducted the behavioral observations. During handling mice were accustomed to a polycarbonate clear mouse handling tube (Plexx BV, Elst, the Netherlands) according to the protocol from Gouveia & Hurst [43]. One handling tube was used per room to habituate the mice. The tube was cleaned with water and a damp tissue between animals. In addition, mice were picked up at the tail base to habituate them to the blood collection procedure for corticosterone measurements (see *2.4 Experimental protocol and blood sampling*).

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 [44][45]. The mHB allows 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 [44]. The paradigm has been described extensively elsewhere [45] and is only briefly explained here. The apparatus consists of a gray PVC opaque box (100 × 50 × 50 cm) with a board made of the same material (60 × 20 × 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 × 15 cm) and 2 squares (20 × 20 cm). The periphery was illuminated with red light (1–5 lux) and functioned as the protected area. In contrast, the central board was illuminated by an additional stage light (120 lux) in order to increase the aversive nature of the central (unprotected) area.

2.4. Experimental protocol and blood sampling

Mice were behaviorally characterized through repeated exposure to the mHB. Testing occurred between 09:30 AM and 2 PM, during the active phase of the animals. Each batch was tested on 4 consecutive testing days within a single week. All mice were tested individually for a total of 4 consecutive trials. Each trial lasted 5 min and test order within batch was randomized across strains.

At the start of the first trial, mice were transferred from the home cage to the mHB using the handling tube, 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 transported back to their home cage using the handling tube and the mHB was 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. Inter-observer reliability was

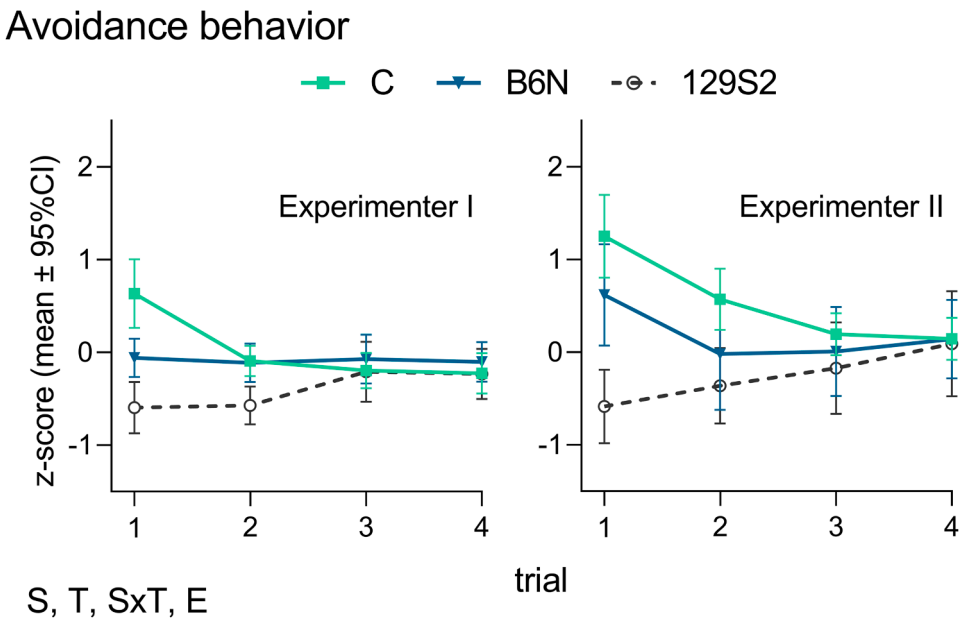

Fig. 1. Display of avoidance behavior across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T a significant main effect of trial; SxT indicates a significant interaction between strain and trial; E denotes a significant main effect of experimenter.

established prior to the start of the study at a moderate to good level [46] with an average Cohen's $\kappa = 0.74$ (range 0.67–0.84) over an average percentage agreement of 94% (range 89.6–97.14). Intra-observer reliability was established at a good level for both experimenters (Experimenter I, average Cohen's $\kappa = 0.83$, range 0.78–0.87; Experimenter II: average Cohen's $\kappa = 0.85$, range 0.70-0.95) over average percentage agreements of 93.35% (range 92.18–95.19) and 97.28% (range = 95.94–98.6) respectively.

Circulating plasma corticosterone levels (pCORT) were assessed for each individual at three sampling moments. The first blood sample was taken one week prior to the behavioral test (7 days ± 1). The second sample was taken directly after behavioral testing, approximately 30 min after the first mHB trial. The third sample was collected one week after behavioral testing (7 days ± 1).

For each individual mouse, it was ensured that all three samples were collected on approximately the same time of day to avoid fluctuation of pCORT due to circadian rhythm [47].

In order to determine baseline pCORT levels in rodents for the first and the last blood sample, the average time from picking up the home cage from the home shelf to finishing the blood collection was recorded (first sample: 120.8 ± 43.4, range 58 – 319; third sample: 127.1 ± 33, range 61 - 276, time in seconds). With an average collection time of 120 s pCORT levels on both levels were considered baseline/not-affected by handling stress. Blood sampling was conducted dropwise via tail vein incision [49], using a single edge industrial blade (GEM®: SPI Supplies, West Chester, PA, USA) by experienced technicians who were not involved in behavioral observations. Sampling occurred in a separate room, mice were transported to this location in their home cage, which was covered with a blanket because the corridor between the two rooms was not under a reversed light-regime.

Blood drops were collected in pre-chilled EDTA coated Microvette® CB300 capillaries (Sarstedt, Nümbrecht, Germany) and stored on ice until plasma was collected by centrifuging the capillary tubes for 30 min on 4 C and 3000 rpm (diameter of the rotor: 17 cm) in a centrifuge (IEC Microlite/Microlite RF®: Thermo Electron Cooperation; West Sussex, UK). Next, plasma (10–20 μl) was pipetted into microtubes and stored at −26 C until further analysis.

2.5. Corticosterone response

Blood plasma corticosterone levels were determined by radioimmunoassay (RIA) according to the manufacturers' protocol of the Corticosterone Double Antibody Kit. Blood samples were coded by number and analyzed in a randomized sequence on the level of individual mouse, so samples from one individual were kept together at all times. Due to technical problems during sampling, or during the laboratory assay there were missing samples (in total $n = 17$, of $n = 17$ individuals).

2.6. Behavioral variables

Behavioral patterns were assessed by scoring behaviors listed in supplementary Table S2. These behaviors were scored as separate variables during testing. However, previous research has shown that the separate behavioral variables scored in the mHB can be reliably allocated to five behavioral dimensions: avoidance behavior, arousal, risk assessment, exploration and locomotion [50][51,52]. In a previous study the identified clusters were composed of these five dimensions [19]. Therefore these same dimensions were again included in the present study.

The separate variables were summarized to their corresponding dimension by using the method of integrated behavioral z-scoring. This method was first proposed by Guilloux et al. [53] and further extended by Labots et al. [52] as a method for behavioral phenotyping in mice. The exact procedure is described in detail elsewhere [19][52] and will therefore only be described briefly here.

In short, behavioral variables that measured different aspects (or different units) of the same behavioral dimension were normalized and combined to a single score representing that particular behavioral dimension or motivational system. Normalization was done by z-score transformation, which measures the amount of standard deviations each observation is above or below the mean of a reference group [52]. The transformed separate variables were averaged within each behavioral dimension. In the present experiment we used the pooled data (across all strains) as the reference group, as suggested by Labots et al. [52]. Supplementary Table S2 presents an overview of all included variables, per behavioral dimension.

2.7. Statistical analyses

2.7.1. Missing values and outliers

The total number of individuals included for statistical analysis per strain was C ($n = 40$), B6N ($n = 39$) and 129S2 ($n = 38$). One B6N mouse was excluded from further analysis due to a procedural error during data collection. Observations that were incomplete (shorter than the 5 min trial length) were labeled as missing value ($n = 4$ trials, of 4 individuals). Furthermore, five trials were identified as influential in the behavioral dimension locomotion using Cook's distance, a commonly used estimate of influential data points in regression analysis [54]. These trials came from two 129S2 animals. One individual had not displayed any locomotion on the first two trials, resulting in the maximum value for the latency to display the first line crossing in these trials (= 300 s, the length of the trial). In addition, this individual displayed a high latency to the first line crossing on trial 3 (> 200 s). The second individual did not display any locomotion on trial 2, and displayed a latency > 200 s to its first line crossing on trial 3. These individuals were not associated with obvious signs of impaired health/wellbeing (as indicated by regular health checks) and were retained for analysis.

2.7.2. Analysis general

All analyses were conducted with R version 3.5.1 in R-Studio [55]. Generalized linear mixed models (GLMMs) were run at several stages of analysis, using the packages 'nlme' [56] and 'glmmTMB' [57]. The specifics of the models used at each stage are described in the subsections below.

At all stages, model assumptions were assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots [58][59]. Heteroscedasticity was avoided using the 'varIdent' variance structure transformation from the 'nlme' package when needed (or its glmmTMB-equivalent). This particular transformation allowed different residual spread for each level of the categorical variables in our model [59]. In addition, all models were run with an autoregressive correlation structure for continuous time covariates (corCAR1).

Main and interaction effects from all linear mixed models (LMMs) were derived using *F*-tests with corresponding *P*-value ($P < 0.05$). Statistical significance of random effects were computed by means of likelihood ratio tests, and reported as Chi Square values. Main and interaction effects of analyses using the package 'glmmTMB' were reported with Wald Chi Square tests, as this package does not (yet) allow extraction of *F*-statistics for testing.

Pairwise comparisons were conducted using the package 'emmeans' [60] 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 [61]. Supplementary Table S10 presents an overview of the adjusted α-value for each comparison. 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'. The guidelines provided by Wahlsten [62] 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, $|d| > 1.5$.

2.7.3. Strain differences (behavior and corticosterone)

GLMM's analyzed strain differences on each behavioral dimension using a 3 (strain) x 2 (experimenter) x 4 (trial) mixed factorial design. Strain, experimenter and trial were included as fixed predictors, as well as their two- and three-way interactions. Individual mouse (ID), slope (trial nested in ID), batch and test order were included as random effects [63].

The variables avoidance behavior, arousal, exploration and locomotion were analyzed using the package 'nlme'. The variable risk assessment was analyzed with the package 'glmmTMB' because the distribution of residuals was zero-inflated. Avoidance behavior was logarithmically transformed and locomotion rank transformed to achieve normality of the residuals. Avoidance behavior included a variance function ('varIdent') for strain (allowing different residual spread between strains) to avoid heteroscedasticity. The variables risk assessment, exploration and locomotion included the same variance function for 'trial' within 'strain' (allowing different residual spread on each trial, for each strain).

Furthermore, the variable pCORT was z-transformed and pCORT levels were analyzed with a generalized least squares (gls) model using a 3 (strain) x 4 (technician) x 3 (sampling moment) mixed factorial design. Strain, sampling moment, technician and the interaction between strain and sampling moment were included as fixed predictors. Day of test was included as a covariate. Individual mouse (ID), slope, batch and test order were initially included as random factors but removed from the model because the model without random factors gave the best fit (as determined by the AIC-criterion). pCORT (nmol/L) was logarithmically transformed to achieve normality of the residuals and the variance function 'varIdent' was applied to allow different residual spread on the three samples within each strain. Detailed results of each explanatory variable, for each behavioral dimension and for pCORT, are provided in supplementary Table S3. Significant main and/or interaction effects were followed up by *post hoc* tests. Detailed results of all *post hoc* comparisons for each dimension and for pCORT are provided in supplementary Tables S4, S5 and S6.

2.7.4. Clustering procedure

Instead of conducting the clustering procedure with the integrated z-scores of the behavioral dimensions and the z-score of pCORT, we used residual values of these z-scores for this part of the analysis. This was done as a means to control for confounding effects of strain, experimenter, batch and test order during assessment of the occurrence of subgroups of individuals that follow a similar behavioral response across trials.

Standardized Pearson residuals of the z-scores were obtained via additional LMMs using a 3 (strain) x 2 (experimenter) factorial design (behavior) or a 3 (strain) x 4 (technician) factorial design (pCORT). The factor 'trial' or 'sampling moment' was intentionally left out of the model because we wanted to maintain this information in the residuals so that we could assess the change in behavior over trials, or in pCORT over sampling moments.

For each behavioral dimension and for pCORT, strain and experimenter/technician were included as fixed factors, with individual mouse (ID), batch and test order as random factors. Avoidance behavior and pCORT were logarithmically transformed to achieve normality of the residuals. Furthermore, avoidance behavior, exploration, locomotion and pCORT included a variance function for 'strain', allowing different residual spread between strains to avoid heteroscedasticity.

The resulting standardized Pearson residual integrated z-scores were subsequently analyzed with a *k*-means clustering procedure using the package 'kml3d' [64]. The settings and rationale for using this particular package have been described in our previous study in detail [19]. The settings used in the present analyses are identical to those specified in [19], with the exception of the distance metric used for clustering. In the present analyses, the Fréchet distance was used because this metric is particularly sensitive for longitudinal data, whereas Euclidean distance was used in our previous study [19].

Six response trajectories were included for each individual mouse: avoidance behavior, risk assessment, arousal, exploration, locomotion and pCORT. These were clustered simultaneously to explore the occurrence of homogeneous groups of mice that followed the same response on all behavioral dimensions. Prior to analysis the gap statistic was applied to evaluate whether the trajectories were perhaps best represented by a single cluster, using the package 'cluster' [65]. 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 [66], 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 number of clusters was selected using the approach of Clustering Validity Indices (CVI's) [67], which was adjusted by Wahl et al. [68]. All details of this procedure are described in [19].

2.7.5. Cluster differences (behavior and corticosterone)

GLMM's analyzed cluster differences on each behavioral dimension using a 2 (cluster) x 4 (trial) mixed factorial design. A LMM analyzed cluster differences in pCORT levels using a 2 (cluster) x 3 (sampling moment) mixed factorial design. In all models cluster and trial/sampling moment were included as fixed predictors, while individual mouse and sampling time (nested in ID) were included as random factors.

The variables avoidance behavior, arousal, exploration, locomotion and pCORT were analyzed using the package 'nlme' [56]. The variable risk assessment was analyzed with the package 'glmmTMB' [57] because the distribution of residuals was zero-inflated.

Locomotion was rank transformed and a square root transformation was applied on pCORT to achieve normality of the residuals. The models for arousal and locomotion included a variance function ('varIdent') for cluster, allowing for differential residual spread between clusters to avoid heteroscedasticity. The models for avoidance behavior, exploration and pCORT included the same variance function for 'trial' (or 'sampling moment' for pCORT) within 'cluster' (allowing differential residuals spread on each trial/sampling moment, within each cluster).

Detailed results of each explanatory variable, for each behavioral dimension and for pCORT, are provided in supplementary Table S7. Significant main and/or interaction effects were followed up by *post hoc* tests. Detailed results of all *post hoc* comparisons for each dimension are provided in supplementary Tables S8 and S9.

2.7.6. Cluster stability

Stability of the clusters was assessed by a bootstrapping procedure in which 200 random samples (of $n = 117$) were drawn from the dataset with replacement (meaning a particular individual could occur multiple times in one sample). If clusters are stable, *kml3d* cluster analyses on all 200 samples should reveal similar cluster structures [69]. Similarity in cluster composition between the bootstrapping samples and the originally obtained clusters was determined by the Jaccard similarity index: For each individual mouse, the number of times (out of 200 bootstrap samples) it belonged to the same cluster as in the original cluster analysis was determined according to 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.

3. Results

3.1. Strain differences (behavior and pCORT)

Generalized linear mixed models (GLMM's) analyzed strain

differences on each behavioral dimension using a 3 (strain) x 2 (experimenter) x 4 (trial) mixed factorial design.

Avoidance behavior trajectories significantly differed between strains (strain effect: $F_{(2, 111)} = 17.44$, $P < 0.0001$; trial effect: $F_{(3, 329)} = 11.04$, $P < 0.0001$; strain x trial interaction: $F_{(6, 329)} = 8.51$, $P < 0.0001$; Fig. 1, supplementary Table S3). *Post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that C decreased ($P < 0.0001$, *very large* effect size, $d = 3.312$, 95%CI [2.228, 4.396]), while 129S2 increased avoidance behavior between trial 1 and trial 4 ($P = 0.0015$, *very large* effect size, $d = -1.670$, 95%CI [−2.712, −0.628]). B6N did not display a significant change across trials (supplementary Table S4).

Further *post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that strains differed in onset levels of avoidance behavior, measured as mean avoidance on trial 1. C mice displayed higher onset levels of avoidance than B6N ($P = 0.0004$, *very large* effect size, $d = 2.372$, 95%CI [1.045, 3.699]) and 129S2 ($P < 0.0001$, *very large* effect size, $d = 5.595$, 95%CI [4.203, 6.987]). Furthermore, 129S2 displayed lower onset levels of avoidance behavior compared to B6N ($P < 0.0001$, *very large* effect size, $d = 3.223$, 95%CI [2.044, 4.402]), supplementary Table S5). The difference in avoidance remained significant on trial 2 between C and 129S2 (adjusted $\alpha = 0.025321$; $P < 0.0001$, *very large* effect size, $d = 2.666$, 95%CI [1.533, 3.799]) and between C and B6N ($P = 0.0057$, *large* effect size, $d = 1.439$, 95%CI [1.045, 3.699]). Overall, strain differences disappeared on trials 3 and 4 (supplementary Table S5).

In addition to these strain differences, avoidance behavior scores differed between experimenters ($F_{(1, 111)} = 8.61$, $P = 0.0041$). Experimenter II (Fig. 1, right panel) scored more avoidance behavior than Experimenter I (Fig. 1, left panel, averaged over trials, $P = 0.0160$, supplementary Table S6). The size of this effect however was small ($d = -0.380$, 95%CI [−0.565, −0.196]).

Risk assessment trajectories differed significantly between strains (strain effect: $\chi^2_{(2)} = 186.53$, $P < 0.0001$; trial effect: $\chi^2_{(3)} = 678.71$, $P < 0.0001$; strain x trial interaction: $\chi^2_{(6)} = 175.72$, $P < 0.0001$), Fig. 2, supplementary Table S3. *Post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that all three strains displayed a significant decrease in risk assessment between the first and the last trial (C, $P < 0.0001$; B6N, $P < 0.0001$; 129S2, $P < 0.0001$, supplementary Table S4). *Post hoc* comparisons (adjusted $\alpha = 0.016952/0.025321$) furthermore showed that on the first three trials, estimates of mean risk assessment were significantly lower for B6N than for C (respectively $P < 0.0001$; $P < 0.0001$; $P = 0.0014$) and lower for B6N than for 129S2 (respectively $P < 0.0001$; $P < 0.0001$; $P < 0.0001$; supplementary Table S5). C and 129S2 did not differ significantly on any of the four trials (supplementary Table S5).

In addition, risk assessment trajectories differed between experimenters (experimenter x trial interaction: $\chi^2_{(3)} = 23.20$, $P = 0.0001$, supplementary Table S3). Scored risk assessment levels were significantly higher on trial 1 (adjusted $\alpha = 0.025321$; $P < 0.0001$, supplementary Table S4) for Experimenter I (Fig. 2, left panel) than for Experimenter II (Fig. 2, right panel). Experimenters did not differ in scored risk assessment behavior on the remaining trials (supplementary Table S6). Overall, the effect size for this experimenter effect was negligible ($d = 0.128$, 95%CI [−0.055, 0.311]).

Arousal trajectories differed between strains (strain effect: $F_{(2, 111)} = 3.49$, $P = 0.0339$; trial effect: $F_{(3, 329)} = 20.66$, $P < 0.0001$; strain x trial interaction: $F_{(6, 329)} = 3.57$, $P = 0.0019$; Fig. 3, supplementary Table S3). *Post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that all three strains significantly increased arousal between the first and the last trial (C, $P = 0.0015$, *moderate* effect size, $d = -0.784$, 95%CI [−1.272, −0.298]; B6N, $P = 0.0006$, *moderate* effect size, $d = -0.888$, 95%CI [−1.397, −0.379]; 129S2, $P < 0.0001$, *very large* effect size, $d = -1.654$, 95%CI [−2.169, −1.140]; supplementary Table S4).

Arousal levels however were highly similar between strains, as post hoc comparisons only showed a significant difference in arousal on trial 2, with higher levels of arousal in B6N compared to 129S2 (adjusted $\alpha = 0.025321$; $P = 0.0002$, moderate effect size, $d = 0.931$, 95%CI [0.436, 1.426]). On the remaining trials, arousal did not differ significantly between strains (supplementary Table S5).

Finally, scored levels of arousal differed between experimenters (experimenter effect: $F_{(1, 111)} = 9.15$, $P = 0.0031$, Fig. 3, supplementary Table S3). Arousal scored by Experimenter I (Fig. 3, left panel) was statistically higher than that of Experimenter II (Fig. 3, right panel, $P = 0.0007$, *small* effect size, $d = 0.409$, 95%CI [0.224, 0.594], supplementary Table S6).

Exploration trajectories differed significantly between strains (strain effect: $F_{(2, 111)} = 38.15$, $P < 0.0001$; trial effect: $F_{(3, 329)} = 20.66$, $P < 0.0001$; strain x trial interaction: $F_{(6, 329)} = 3.73$, $P = 0.0113$, Fig. 4 and supplementary Table S3). *Post hoc* comparisons (adjusted $\alpha = 0.016952$)

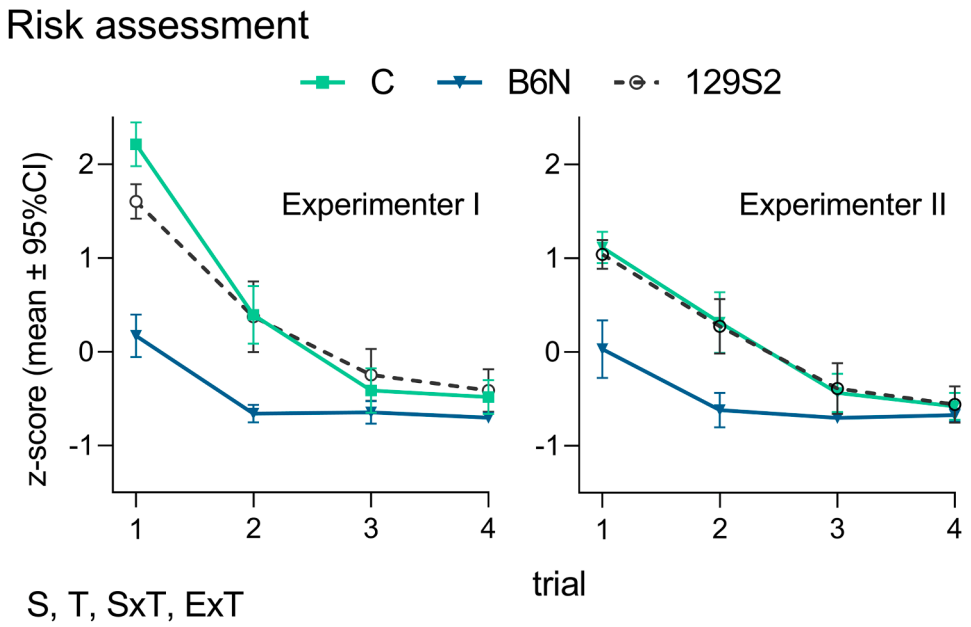

Fig. 2. Display of risk assessment across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a GLMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates a significant main effect of trial; SxT indicates a significant interaction between strain and trial; ExT denotes a significant interaction between experimenter and trial.

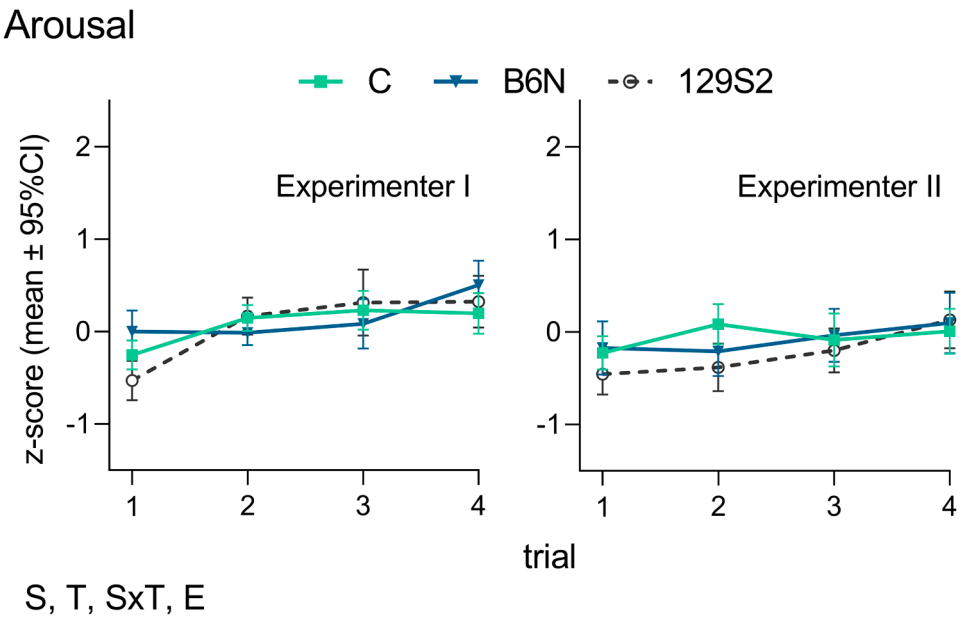

Fig. 3. Display of arousal across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates a significant main effect of trial; SxT indicates a significant interaction between strain and trial; E denotes a significant effect of experimenter.

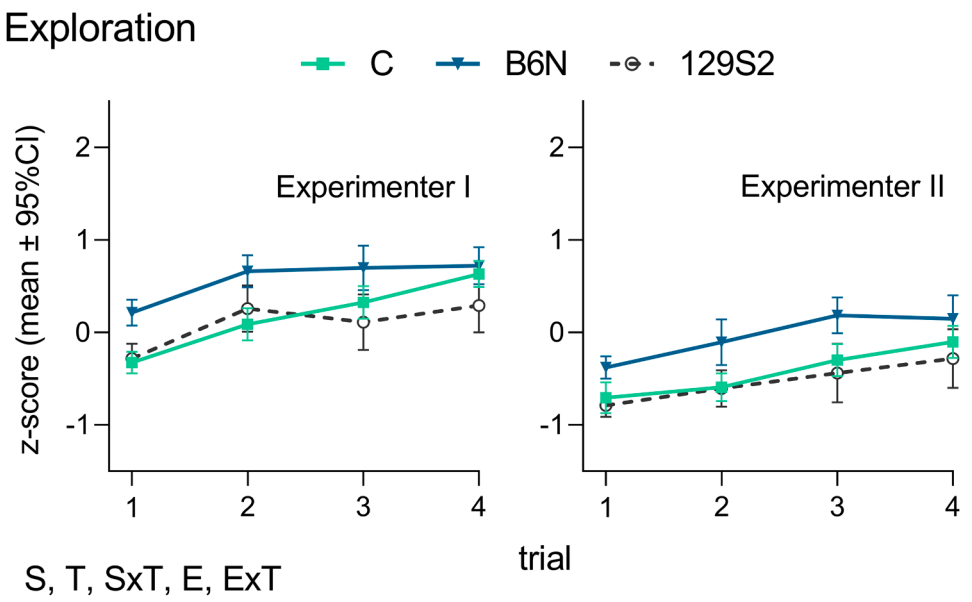

Fig. 4. Display of exploration across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates a significant main effect of trial; SxT indicates a significant interaction between strain and trial; E denotes a significant effect of experimenter; ExT indicates a significant interaction between experimenter and trial.

indicated that all three strains significantly increased exploration between the first and the last trial (C, $P < 0.0001$, *very large* effect size, $d = -2.591$, 95%CI [−3.070, −2.113]; B6N, $P < 0.0001$, *very large* effect size, $d = -1.701$, 95%CI [−2.208, −1.194]; 129S2, $P < 0.0001$, *very large* effect size, $d = -1.787$, 95%CI [−2.432, −1.142], supplementary Table S4).

Post hoc comparisons (adjusted $\alpha = 0.016952/0.025321$) further showed that B6N displayed higher levels of exploration than 129S2 on all four trials, and higher levels of exploration than C mice on the first three trials (supplementary Table S5). Mean exploration did not differ between C and 129S2 mice on any of the trials (supplementary Table S5).

In addition, exploration trajectories also differed between experimenters (experimenter effect: $F_{(1,\ 111)} = 114.95$, $P < 0.0001$; experimenter x trial interaction: $F_{(3,\ 329)} = 5.08$, $P = 0.0019$, Fig. 4, supplementary Table S3). *Post hoc* comparisons (adjusted $\alpha = 0.025321/0.05$) comparing exploration scores between experimenters on each trial showed that observed exploration was significantly higher for Experimenter I than for Experimenter II on all four trials (Fig. 4, supplementary Table S6). These differences were accompanied by very large effect sizes on all trials (supplementary Table S6).

The model for locomotion (rank transformed) showed that the trajectories for locomotion differed significantly between strains (strain effect: $F_{(2,\ 111)} = 183.02$, $P < 0.0001$; trial effect: $F_{(3,\ 329)} = 18.91$, $P < 0.0001$; strain x trial interaction: $F_{(6.\ 329)} = 20.94$, $P < 0.0001$; Fig. 5, supplementary Table S3). *Post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that B6N significantly decreased, while C significantly increased locomotion between trial 1 and trial 4 (B6N, $P < 0.0001$, *very large* effect size, $d = 1.715$, 95%CI [1.375, 2.054]; C, $P < 0.0001$, *moderate* effect size, $d = -0.971$, 95%CI [−1.421, −0.520], supplementary Table S4).

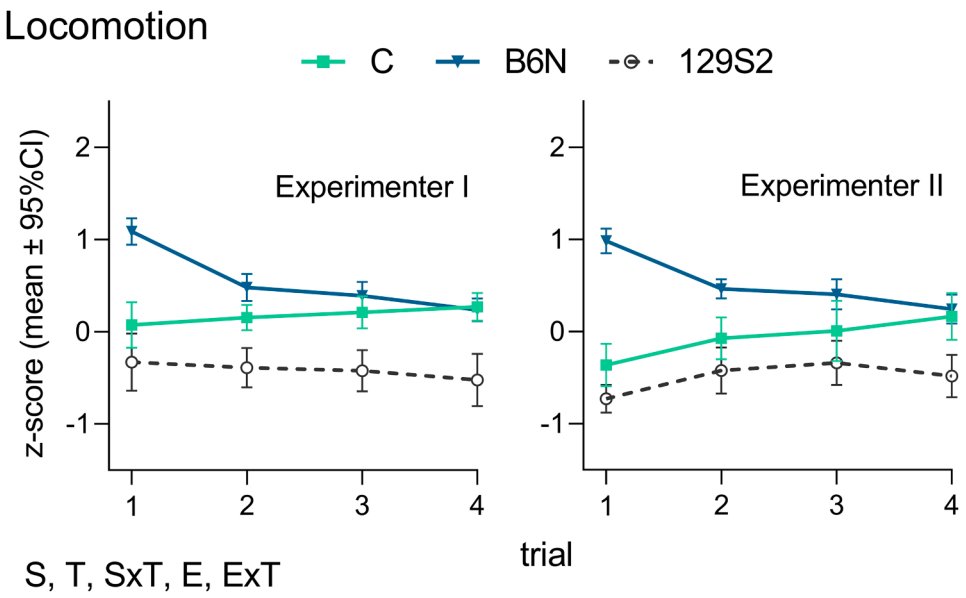

Fig. 5. Display of locomotion across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates a significant main effect of trial; SxT indicates a significant interaction between strain and trial; E denotes a significant effect of experimenter; ExT indicates a significant interaction between experimenter and trial.

129S2 did not display a significant change in locomotion between these trials (supplementary Table S4). Furthermore, locomotion was significantly higher for B6N than for 129S2 on all four trials, and higher than C on the first three trials (supplementary Table S5). In addition, locomotion was also significantly higher for C mice than for 129S2 on all four trials (supplementary Table S5).

Locomotion scores per strain also differed between experimenter (strain x experimenter interaction: $F_{(2,\ 111)} = 3.50$, $P = 0.0336$, supplementary Table S3). *Post hoc* comparisons indicated that overall locomotion scores for C mice were higher Experimenter I than for Experimenter II ($P = 0.0232$, moderate effect size, $d = 0.588$, 95%CI [0.076, 1.100], Fig. 5, supplementary Table S6). Experimenters scores of locomotion were nog significantly different for 129S2 and B6N (supplementary Table S6).

The trajectories of pCORT differed significantly between strains (strain effect: $F_{(2,\ 321)} = 21.81$, $P < 0.0001$; sampling moment effect: $F_{(2,\ 321)} = 263.98$, $P < 0.0001$; strain x sampling moment interaction: $F_{(4,\ 321)} = 4.67$, $P = 0.0011$; Fig. 6, supplementary Table S3).

Post hoc comparisons (adjusted $\alpha = 0.012741$) revealed that pCORT levels were higher on sampling moment 2 than baseline (sampling moment 1) for all three strains (C, $P = 0.0009$, *large* effect size, $d = -1.389$, 95%CI [−2.194, −0.585]; B6N, $P = 0.0002$, *very large* effect size, $d = -1.676$, 95%CI [−2.541, −0.812]; 129S2, $P = 0.0002$, *very large* effect size, $d = -2.460$, 95%CI [−3.701, −1.219]; supplementary Table S4).

Furthermore, *post hoc* comparisons (adjusted $\alpha = 0.012741$) showed that strains did not significantly in baseline pCORT levels (supplementary Table S5). Directly after behavioral test however (sampling moment 2), pCORT was significantly higher for 129S2 mice than for B6N ($P = 0.00281$, *moderate* effect size, $d = -0.979$, 95%CI [−1.694, −0.264]) and there was a suggestive effect of higher pCORT in 129S2 compared to C ($P = 0.0104$, *large* effect size, $d = -1.102$, 95%CI [−1.937, −0.267]; supplementary Table S5).

A week after behavioral test (sampling moment 3), pCORT had decreased significantly in all three strains compared to sampling moment 2 (C, $P = 0.0028$, *large* effect size, $d = 1.069$, 95%CI [0.375, 1.762]; B6N, $P = 0.0003$, *very large* effect size, $d = 1.616$, 95%CI [0.772, 2.460]; 129S2, $P = 0.0002$, *very large* effect size, $d = 2.273$, 95%CI [1.110, 3.435], supplementary Table S4). On this sampling moment however, pCORT levels were significantly lower for B6N than for C ($P = 0.0030$, small effect size, $d = 0.424$, 95%CI [0.151, 0.698]; supplementary Table S5).

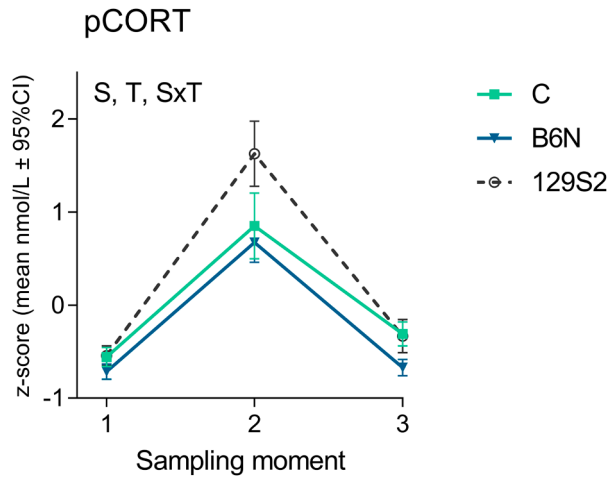

Fig. 6. Blood plasma corticosterone (pCORT) levels in strains C, B6N and 129S2 one week prior to behavioral test (sampling moment 1), directly after behavioral test (sampling moment 2) and one week after behavioral test (sampling moment 3). Results are expressed as z-transformed nmol/L and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates a significant main effect of time (sampling moment); SxT indicates a significant interaction between strain and time (sampling moment).

Thus, in all three strains pCORT levels were significantly higher directly after behavioral test, compared to a week before test, with 129S2 presenting the highest pCORT levels directly after test compared to the other two strains. One week after behavioral testing, pCORT levels had decreased significantly in all strains, with C being the only strain with significantly higher pCORT levels on sampling moment 3, compared to baseline on sampling moment 1 (adjusted $\alpha = 0.01695$; $P = 0.0077$, small effect size, $d = -0.320$, 95%CI [−0.555, −0.086], supplementary Table S4).

3.2. *Cluster analysis*

Standardized Pearson residuals of the integrated z-scores were used for the clustering of the individual trajectories (see Section 2.7.4). Data from all strains was pooled in order to assess individual variation in

habituation responses within and across strains. All five behavioral dimensions and pCORT were taken into account simultaneously. As such, six response trajectories were included for each individual mouse: avoidance behavior, risk assessment, exploration, locomotion, arousal and pCORT. The optimal partitioning of the data yielded two clusters. Table 1 presents cluster size and distribution of strains across clusters. The mice were more or less evenly distributed across clusters, with 53.8% of mice (n = 63) grouping together in cluster A while the remaining 46.2% (n = 54) fell in cluster B.

Cluster A and cluster B both consisted of individuals from all three strains, but the distribution of strains differed between clusters. The majority of C (82.5%, n = 33 out of 40) fell in cluster A, while the majority of 129S2 (76.3%, n = 29 out of 38) grouped together in cluster B. B6N mice were divided over clusters A (53.8%, n = 21) and B (46.2%, n = 18). Because of the marked experimenter effects in the strain analyses, the distribution of mice within clusters are presented for each experimenter separately.

3.3. Cluster differences (behavior and pCORT)

Fig. 7 presents the trajectories of clusters A and B on each behavioral dimension as well as on pCORT. The trajectories of the clusters differed on all behavioral dimensions, except for risk assessment. This behavior decreased as trials progressed, regardless of cluster (trial effect: $\chi^2_{(3)}$ = 417.49, $P < 0.0001$, Fig. 7, supplementary Table S7).

Avoidance behavior trajectories differed significantly between clusters (trial effect: $F_{(3,\ 341)}$= 10.27, $P < 0.0001$; cluster x trial interaction: $F_{(3,\ 341)}$= 45.32, $P < 0.0001$), Fig. 7). *Post hoc* comparisons (adjusted α = 0.02532) showed that cluster A decreased ($P < 0.0001$, *very large* effect size, d = 1.558, 95%CI [1.247, 1.869]) between the first and the last trial. In contrast, cluster B increased avoidance behavior between trial 1 and trial 4 ($P < 0.0001$, large effect size, d = −1.000, 95%CI [−1.382, −0.618]), supplementary Table S8. Further *post hoc* comparisons (adjusted α = 0.02532/0.05) showed that avoidance behavior differed significantly between clusters on all trials apart from trial 3 (trial 1, $P <$ 0.0001, *very large* effect size, d = 1.860, 95%CI [1.272, 2.448]; trial 2, P = 0.0019, *medium* effect size, d = 0.690, 95%CI [0.250, 1.130]; trial 4, P = 0.0014, *medium* effect size, d = −0.698, 95%CI [−1.129, −0.267], Fig. 7, supplementary Table S9).

The trajectories of arousal also differed across trials between clusters (trial effect: $F_{(3,\ 341)}$= 18.41, $P < 0.0001$; cluster x trial interaction: $F_{(3,\ 341)}$= 8.59, $P < 0.0001$; Fig. 7). *Post hoc* comparisons (adjusted α = 0.02532) showed that both clusters increased arousal between the first and the last trial (cluster A, P = 0.0010, *medium* effect size, d = −0.641, 95%CI [−1.026, −0.256]; cluster B, $P < 0.0001$, *very large* effect size, d = −1.875, 95%CI [−2.420, −1.329], supplementary Table S8). Cluster B displayed higher levels of arousal on the last two trials (adjusted α = 0.02532/0.05), indicating that arousal increased more pronounced in this cluster (trial 3, P = 0.0008, *medium* effect size, d = −0.849, 95%CI [−1.350, −0.348]; trial 4, P = 0.0020, *medium* effect size, d = −0.834, 95%CI [−1.367, −0.300], Fig. 7, supplementary Table S9).

Furthermore, clusters differed significantly with respect to activity

Table 1
Top row: Cluster size and proportion of total population per cluster. Bottom rows: Distribution of mice across strains, experimenters and clusters (n and proportion).

Cluster size (n) and proportion of total n per cluster						
	Cluster A			Cluster B		
n total = 117	n = 63 *(53.8%)*			n = 54 *(46.2%)*		
Distribution of strains *within* clusters and per experimenter						
	Cluster A			Cluster B		
	Exp. I	Exp. II	Total	Exp. I	Exp. II	Total
Strain	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
C	16 (48.5)	17 (51.5)	33 (52.4)	4 (57.1)	3 (42.9)	7 (13.0)
B6N	12 (57.1)	9 (42.9)	21 (33.3)	12 (66.7)	6 (33.3)	18 (33.3)
129S2	3 (33.3)	6 (67.7)	9 (*14.3*)	16 (55.2)	13 (44.8)	29 (53.7)

related dimensions. Locomotion (rank transformed) was significantly higher in cluster B compared to cluster A regardless of trial (cluster effect: $F_{(3,\ 341)}$ = 11.35, P = 0.0010; Fig. 7, supplementary Table S7).

Exploration trajectories differed between clusters (trial effect: $F_{(3,\ 341)}$ = 59.73, $P < 0.0001$; interaction cluster x trial: $F_{(3,\ 341)}$ = 8.59, $P <$ 0.0001; Fig. 7, supplementary Table S7). *Post hoc* comparisons (adjusted α = 0.02532) however indicated that both clusters increased exploration between trial 1 and trial 4 (cluster A, $P < 0.0001$, *very large* effect size, d = −2.949, 95%CI [−2.905, −2.083]; cluster B, P = 0.0001, *medium* effect size, d = −0.935, 95%CI [−1.348, −0.539]; supplementary Table S8). The significantly higher exploration levels in cluster A compared to B on trials 3 and 4 however suggest the increase in exploration was more pronounced in cluster A (trial 3, P = 0.0458; *medium* effect size, d = 0.647, 95%CI [0.007, 1.287; trial 4, P = 0.0023, *large* effect size, d = 1.065, 95%CI [0.374, 1.757]; Fig. 7, supplementary Table S9).

Finally, pCORT did not differ between clusters. A LMM analyzing cluster differences across sampling time for pCORT only revealed a significant effect for sampling moment ($F_{(2,\ 213)}$ = 224.8, $P < 0.0001$; Fig. 7, supplementary Table S7). Post hoc comparisons (adjusted α = 0.02532) showed that regardless of cluster, pCORT levels were higher directly after behavioral testing (sampling moment 2) than a week prior to testing ($P < 0.0001$, *very large* effect size, d = −8.707, 95%CI [−10.720, −7.141]) and a week after behavioral testing ($P < 0.0001$, *very large* effect size, d = 7.905, 95%CI [6.400, 9.409]). Also, pCORT was significantly higher in sample 3 than on sample 1 (P = 0.0001, *medium* effect size, d = −0.801, 95%CI [−1.220, −0.387]; Fig. 7, supplementary Table S8).

To summarize, mice in cluster A decreased avoidance behavior, while exploration increased more pronounced, and overall levels of locomotion were higher compared to cluster B, indicating a successful habituation of initial high anxiety responses. Initial levels of avoidance behavior were low in cluster B, but this behavior increased across trials. At the same time, increase in exploration was less pronounced and mice in this cluster displayed lower levels of locomotor activity.

3.4. Relative weight of dimensions on cluster partitioning

The obtained clusters were based on simultaneous clustering of all five behavioral dimensions and pCORT. However, cluster differences were more pronounced on some variables than on others, with significant differences between clusters in avoidance behavior, arousal, exploration and locomotion, but not in risk assessment and pCORT levels.

In order to assess the relative impact of each variable on this partitioning, we conducted additional cluster analyses, each time leaving one of these six dimensions out. Pearson Chi square tests showed that cluster size in any of these analyses did not significantly differ from cluster size in the original cluster analysis (Table 2).

The Jaccard similarity index subsequently indicated how many individual mice were retained in the same cluster as in the original cluster analysis, after excluding a certain behavioral dimension. As can be seen in Table 2, only 53% of the mice retained their cluster after omitting avoidance behavior, while omitting any of the other dimensions hardly affected cluster membership for individual mice (Jaccard indices > 0.90, Table 2). Thus, avoidance behavior appeared most dominant in partitioning of the clusters.

3.5. Cluster stability

Fig. 8 depicts the mean trajectory of all bootstrap samples (black dashed line) against the trajectory belonging to the original cluster (cluster A, orange; cluster B, blue), as well as the 200 trajectories of the bootstrap samples (gray), for each cluster, on each dimension. For cluster A, the average Jaccard similarity index was 0.64, meaning that on average, an individual mouse belonged to cluster A in 64% of the

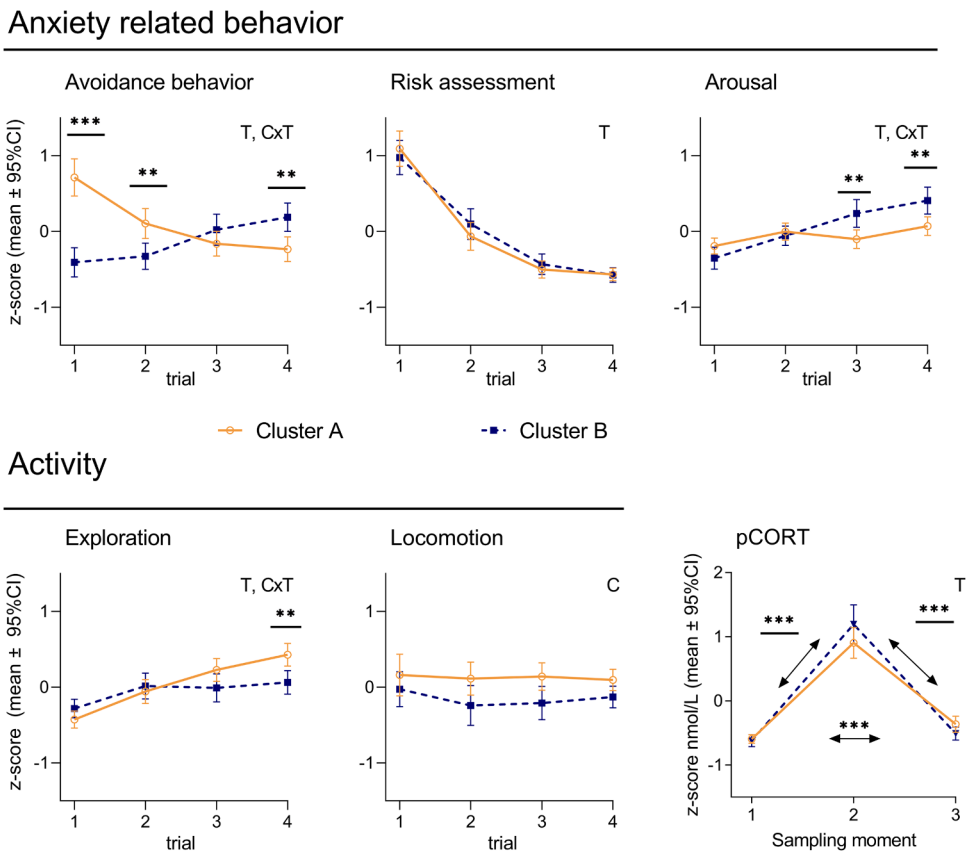

Fig. 7. Differences between clusters on each behavioral dimension, and corticosterone levels. Behavior expressed as integrated behavioral z-scores for behavioral dimensions, and the z-score nmol/L for pCORT. Results are presented as means with 95% CI. Effects were significant in LMMs at $P < 0.05$. C indicates a significant main effect of cluster; T indicates a significant main effect of trial (for behavioral dimension) or sampling time (for corticosterone); C x T indicates a significant interaction between cluster and trial/time. *Behavioral dimensions*: Significant differences in *post hoc* comparisons between clusters on trials 1 and 4 (adjusted $\alpha = 0.025321$) are indicated with ** = $0.000050 \geq P < 0.00501$, *** = $P < 0.00050$. Significant comparisons between clusters on trials 2 and 3 ($\alpha = 0.05$) are indicated with * = $0.01 \geq P < 0.05$; ** = $0.001 \geq P < 0.01$. Corticosterone: significant *post hoc* differences between sampling moments (adjusted $\alpha = 0.025321$) indicated by *** = $P < 0.00050$.

Table 2
Overview of number of mice per cluster when omitting one of the five behavioral dimensions or pCORT.

	All included	Excluded					
		AVO[a]	RA[a]	AR[a]	EXPL[a]	LOC[a]	pCORT
Cluster A (*n*)	65	64	57	59	63	56	53
Cluster B (*n*)	52	53	60	58	54	61	64
P-values (Pearson Chisq)	–	0.694	0.794	1.000	0.794	0.695	0.433
Jaccard Index	–	0.53	0.97	0.99	0.96	0.96	0.92

[a]AVO = avoidance behavior; RA = risk assessment; AR = arousal; EXPL = exploration; LOC = locomotion.

bootstrap samples. The average Jaccard similarity index for cluster B was 0.53.

3.6. Cluster differences within strain (behavior and corticosterone)

Finally, each cluster consisted of individuals of all three strains. These strains showed quite distinct behavioral profiles, of differential adaptive value. Fig. 9 therefore provides a visual representation of how the identified clusters were expressed within each strain separately.

LMM's analyzed cluster differences for each dimension within each strain using a 2 (cluster) x 4 (trial) mixed factorial design. Cluster, trial and their interaction were included as fixed predictors, and individual mouse as random factor. To avoid repetition of statistical results, the specification of cluster differences within each strain in the section below is purely descriptive. Detailed results of all analyses are presented in supplementary tables S11 and S12. For a quick reference, significant main and interaction effects, and any significant *post hoc* tests comparing cluster differences per trial are depicted in Fig. 9.

3.6.1. C

On average, C mice successfully habituated to the test. Within this strain however, subgroups of mice differed significantly in avoidance behavior and arousal (Fig. 9, Supplementary Table S11).

The majority of C ($n = 33$) grouped together in cluster A (Table 2). The behavioral profile of this cluster was indeed highly similar to the average C response: initial high levels of avoidance behavior and risk assessment that significantly decreased, while arousal, exploration and locomotion significantly increased across trials (Fig. 9, top row; Supplementary Table S12).

A small subgroup of C (cluster B, $n = 7$) however, displayed significantly lower levels of avoidance behavior than cluster A on the first trial, and higher levels of avoidance behavior on trial 4 (Fig. 9, top row; Supplementary Table S12). In addition, avoidance behavior remained stable across trials (Supplementary Table S12). This subgroup also displayed a more rapid increase in arousal compared to their counterparts in cluster A, with significantly higher levels of arousal on the last two trials (Fig. 9, top row, Supplementary Table S12).

3.6.2. B6N

On average, B6N were characterized by high activity, and low anxiety levels. This strain was distributed almost equally across clusters, with $n = 21$ individuals in cluster A, and $n = 18$ individuals in cluster B (Table 2). Within B6N, these clusters differed significantly on avoidance behavior, arousal, exploration and locomotion (Fig. 9, middle row; Supplementary Table S11).

On average, avoidance behavior remained stable across trials in B6N.

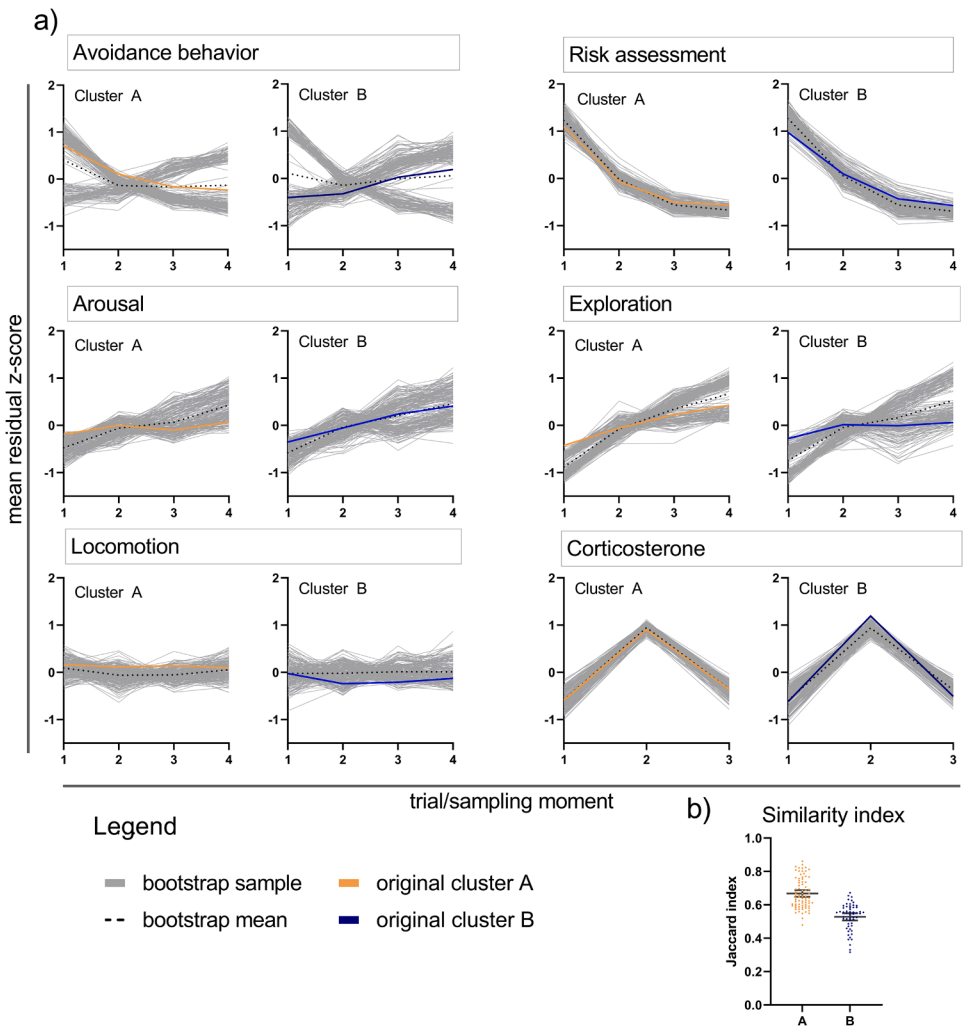

Fig. 8. Results of the bootstrapping procedure for each cluster, on each behavioral dimension. Results are presented as mean residual z-scores, and depict the trajectory of the original cluster (cluster A, orange; cluster B, blue) in relation to the average of all 200 bootstrap samples (black dashed line), against all 200 trajectories of the bootstrapping procedure (gray). **(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).

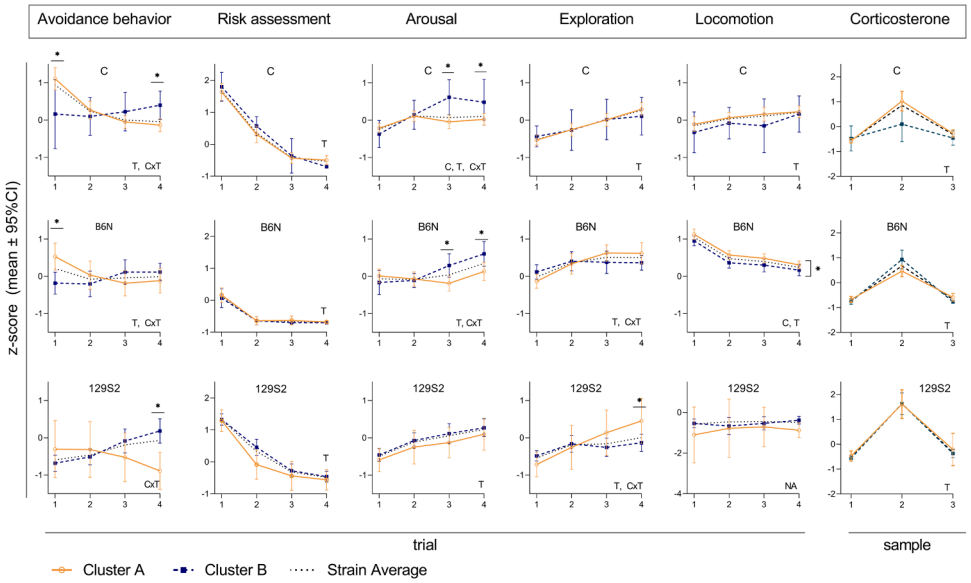

Fig. 9. Differences between clusters within strains (C, top row; B6N, middle row; 129S2, bottom row) on each behavioral dimension, and corticosterone levels. Behavior expressed as integrated behavioral z-scores for behavioral dimensions, and the z-score nmol/L for pCORT. Results are presented as means with 95%CI. Effects were significant in LMMs at $P < 0.05$. C indicates a significant main effect of cluster; T indicates a significant main effect of trial (for behavioral dimension) or sampling time (for corticosterone); C x T indicates a significant interaction between cluster and trial/time. *Behavioral dimensions*: Significant differences in *post hoc* comparisons between clusters on trials 1 and 4 (adjusted $\alpha = 0.025321$) are indicated with * = $P < 0.025321$. Significant comparisons between clusters on trials 2 and 3 ($\alpha = 0.05$) are indicated with * $P < 0.05$.

The identified clusters however displayed contrasting patterns of avoidance behavior across trials: a significant decrease in cluster A, and a significant increase in cluster B between the first and the last trial (Fig. 9, middle row; Supplementary Table S12), with significantly higher levels of avoidance behavior on trial 1 for B6N mice in cluster A (Supplementary Table S12). Arousal increased in both clusters for B6N, but

arousal was higher on the last two trials in cluster B compared to cluster A (Fig. 9, middle row; Supplementary Table S12). Exploration did not differ between clusters on any of the trials, and overall locomotion was higher in B6N mice in cluster A than in cluster B (Supplementary Table S12).

3.6.3. 129S2

The average behavioral response of 129S2 mice was indicative of a sensitization of anxiety responses, which was primarily indicated by initial low levels of avoidance behavior that increased over trials (Fig. 2). At the same time overall levels of locomotor activity were lowest of all strains (Fig. 6).

The majority of 129S2 grouped together in cluster B ($n = 28$, 63.2%), while the remaining individuals fell in cluster A ($n = 9$, 36.8%). Within this strain, the two clusters differed significantly on avoidance behavior and exploration (Fig. 9, bottom row; Supplementary Table S11).

Avoidance behavior in cluster B increased significantly, while this behavior decreased in 129S2 mice in cluster A, with significantly lower avoidance behavior on trial 4 in cluster A than in cluster B (Fig. 9, bottom row; Supplementary Table S12).

Exploration trajectories differed between clusters, with significantly higher exploration on trial 4 for 129S2 mice in cluster A (Fig. 9, bottom row; Supplementary Table S12).

4. Discussion

We showed that behavioral responses to novelty, measured as habituation or sensitization of anxiety-related behavior in the mHB, differ between individuals in three commonly used mouse inbred strains. The behavioral profiles of the identified clusters were largely similar to the habituation and sensitization profile previously identified by van der Goot et al. [19].

As noted in the introduction, we asked whether these previously identified clusters were perhaps (part) result of the variation that was inherent to re-analyzing a dataset consisting of multiple experiments. This aggregated dataset varied naturally in factors that are known to affect variability between experiments, such as test location, time of year, experimenter, etcetera [20][21]. The consistency in behavioral profiles between [19] and the present experiment however suggests that this was not the case. As such this study empirically confirms that adaptive capacities may be subject to inter-individual variability in 129 mice.

Also in line with van der Goot et al. [19], the identified response types were multidimensional, with individual mice grouping together across both anxiety and activity related behavioral dimensions. This confirms the notion that anxiety is a complex behavioral construct that is expressed by multiple behavioral dimensions [18,27,71].

At the same time, the differential behavioral profiles were not substantiated with differential endocrine profiles. Glucocorticoid responses have been associated with anxiety-related behavior in mice [28], and have been found to vary greatly between individuals in rodents [29,30,70,72,73].

Previous findings showed that differential behavioral response types may be correlated to differences in endocrine response in inbred strains [32,33,34]. Jakovcevski et al. [32] for example identified sub-populations of B6 exhibiting high and low trait anxiety (measured as the latency to freely enter a novel environment from their home cage) and showed that trait anxiety was positively correlated with pCORT concentrations after exposure to a stressor [32]. This correlation however was highly dependent on the type of stimulus used. The association was only found after individuals were exposed to a severe stressor (exposure to a rat), and not when mice were exposed to a mild stressor, a novel environment [32]. Exposure to novelty, a major characteristic of the paradigm employed in our experiment, is indeed typically regarded as a mild stressor [74] and this may have affected our results.

On a behavioral level however, clustering individual response trajectories yielded two multidimensional subtypes of response (Fig. 7). In line with [19], avoidance exerted the most 'weight' on the partitioning of the clusters, suggesting that this behavioral dimension is an important distinguishing factor of these multidimensional response types (Table 2).

New to the findings by van der Goot et al. [19] was that these behavioral subtypes were not only displayed by 129 mice, but also present in C and a newly included strain, B6N. These three strains are known for their differential innate emotionality [17], and this should be taken into account when interpreting the differential response types.

The highly neophobic [18] but adaptive phenotype that is characteristic for C for example, [11][14–16]was also observed in the present study. On average C mice displayed initial high levels of anxiety-related behavior that decreased rapidly, while initial low levels of activity increased with trials. The individual based analyses demonstrated that the majority of C indeed grouped together in cluster A, which mirrored the adaptive profile of C mice ($n = 33$, 82.5%, Table 1). A small subgroup however deviated from this response with low initial levels of avoidance behavior and a more pronounced increase in arousal, while overall levels of locomotion were lower than their counterparts in cluster B ($n = 7$, 17.5%, Table 1, Fig. 9).

In general, C mice are often characterized as phenotypically robust, showing relatively little within strain variation compared to other inbred strains [75]. At the same time, subtypes in emotional reactivity and sensitivity to stress have been identified previously in these animals [76]. The seemingly less adaptive profile could suggest that such inter-individual variability may also pertain to the adaptive anxiety phenotype that is characteristic for this strain. At the same time, it should be kept in mind that the small number of C mice that deviate from their average strain response may equally represent individuals that are less responsive to the test, and further assessment of these subtypes, for example by means of pharmacological validation, could provide more insight on this.

Next, the non-adaptive phenotype that is characteristic for various sub-strains of 129 mice [11,13–16] was confirmed in the present study. On average avoidance behavior increased in 129S2, while low levels of locomotion remained stable across trials and exploration increased. In addition, average pCORT levels were significantly higher in 129S2 mice directly after behavioral testing compared to C- and B6N mice. This finding in itself was interesting because it has been speculated before, that the lack of habituation in 129S2 mice (i.e. a lack of exploration of the unprotected area in the mHB) might be due to persistent low levels of locomotor activity and not related to anxiety-related characteristics [13].

Substrains of 129S2, including 129S2/SvPasCrl tested here, are indeed known for their reduced activity levels [26,77,78]. That pCORT levels were highest in 129S2 mice however, suggests that exposure to the mHB was indeed perceived as particularly stressful by this strain, and could serve as an indication that the anxiety phenotype of 129S2 can be classified as non-adaptive. At the same time, there are many additional factors that affect corticosterone levels at a given time point (i.e. circadian rhythm [79], intestinal microbiome [80], nutritional status [81]), so further research is necessary to determine whether the elevated corticosterone levels in 129S2 mice were indeed associated with increased anxiety or whether other factors were at play.

In line with [19], the majority of 129S2 mice grouped together in the cluster that mirrored the average behavioral non-adaptive 129S2 anxiety phenotype (Cluster B, $n = 29$, 76.3%, Table 1). Also in line with this study, in a small subgroup of 129S2 mice (cluster A, $n = 9$, 23.7%, Table 1) anxiety-related behavior decreased across trials, suggesting a seemingly more adaptive phenotype (Fig. 7), with lower levels of avoidance behavior on the last trial, lower risk assessment, risk assessment and arousal (all three not significant) and a more pronounced increase in exploration, than their counterparts in cluster B (Fig. 9). This could suggest that not all 129S2 individuals are equally susceptible to the aversive nature of the mHB.

In comparison to C and 129S2, B6 mice are often characterized as a low anxious and highly active strain [22][23][24]. This was confirmed by the average behavioral profile in the current study, in which B6N mice expressed high levels of exploration and locomotor activity and low levels of anxiety-related behavior. More specifically, avoidance behavior remained stable across trials, while exploration increased and locomotion decreased, which corroborates previous assessment of B6 in the mHB [27][44]. pCORT levels have been found to be lower in B6 mice in comparison to C mice when tested in the mHB [82]. Our finding that pCORT levels were lower than in 129S2 and (not significantly) in C mice extend this observation.

Despite this low anxious profile however, B6N were almost evenly distributed across the two clusters with $n = 21$ (53.8%) in cluster A and $n = 18$ (46.2%) in cluster B (Table 1). On average, avoidance behavior did not change across trials in this strain, but the division of B6N mice across the two clusters indicated that in fact avoidance behavior decreased in some individuals (cluster A) while it increased across trials in others (cluster B). This demonstrates that a more in-depth individual characterization of B6 mice in the mHB may reveal contrasting patterns of avoidance behavior that would be overlooked when only focusing on average strain responses.

The fact that on average B6N mice displayed a low anxious, highly active profile, and the fact that further zooming in on these individual responses showed that within B6N, avoidance behavior only differed significantly on the first trial between clusters (Fig. 9), raises the question, whether the identified subtypes in B6N in the mHB should indeed be interpreted as differential anxiety phenotypes.

B6 is often used as a comparator strain when testing for anxiety [83]. Despite its reputation as a low anxious strain however, inter-individual variability in anxiety related behavior has repeatedly been demonstrated [32][33][76][84][85]. In fact, the majority of studies that identified subtypes of emotional reactivity within inbred strains have been conducted with B6-mice. These distinct anxiety-profiles in B6 mice have been found consistent across time and context [84][85], and have even been linked to copy number variation in small nucleolar RNA clusters, suggesting a genetically specified modulator of these differential profiles [85]. At the same time, one could interpret the identified subtypes as indicative of differential activity levels, rather than indicative of differential anxiety profiles. The two clusters indeed differed in exploration and locomotion and inter-individual variability in activity related behavior has been established previously in B6 mice [86]. The existing literature does not provide a definitive explanation in either direction. Further zooming in on these subtypes is necessary to evaluate to what extent these profiles of B6 mice are related to differential activity levels, or whether they also reflect differential anxiety responses.

This latter point in fact holds for all three strains. The identified subtypes not only displayed contrasting patterns of avoidance behavior, but were also characterized by overall differences in locomotor activity. Locomotion is not only associated with general activity, but may also exert a confounding effect on anxiety related behavior [71]. A lack of exploration of a certain area (i.e. avoidance behavior) may however, merely be the result of reduced locomotor activity [13][87] and thus independent from anxiety.

The lower activity levels of the small sub-population of C mice could simply represent individuals that were less responsive to the aversive nature of the mHB. Conversely, the higher activity levels in the small sub-group of 129S2 individuals compared to the majority of 129S2 mice could have affected the observed decrease in avoidance behavior.

Locomotion has repeatedly been dissociated from avoidance behavior by factor analyses on the behavioral variables observed in the mHB [27][51]. Labots et al. [87] however also observed that substrain differences in avoidance behavior disappeared after controlling for horizontal locomotor activity (i.e. the type of locomotor activity that was recorded in the present study), suggesting that locomotion may in fact exert a confounding effect on avoidance behavior in the mHB. In the present study, differences in avoidance behavior between clusters remained intact when including locomotor activity as a covariate in the model (trial effect: $P < 0.0001$; interaction strain x trial: $P < 0.0001$).

The current results unfortunately do not allow for a definitive dissociation between anxiety-related behavior and a potentially confounding effect of locomotion. Further research should therefore first be aimed at ruling out this potentially confounding effect, for example by assessing whether the observed subtypes respond differently to pharmacological treatment, or by combining the currently employed assay with a behavioral test that is less dependent on locomotor activity, such as the physiological anxiety paradigm stress-induced hyperthermia [89].

Further pharmacological validation may also provide more insight in the question whether the identified subtypes indeed reflect two qualitatively differential anxiety responses, or whether these sub-populations represent individuals that were simply were less responsive to the test. As noted above, given the strain specific differences in emotionality, such further assessment is ideally determined within each strain separately.

When further evaluating the potential differential anxiety profiles in these strains, it should furthermore be assessed whether the identified variation is consistent across time and contexts [6][9]. As described earlier, the advantages of individual-based characterization are considered twofold: It may enable the selection of susceptible or (un) responsive individuals from a cohort – provided that sufficient time is allowed before retesting in the same or other paradigms [89] - and could thereby make animal models more representative. Second, identifying subtypes of response could provide a starting point for the exploration of the biological mechanisms underlying these subtypes [6].

These presumed benefits are however highly dependent on the temporal consistency of the behavior of interest [6][9][85]. Consistency of anxiety-related and activity behaviors across time [84][85][86] and contexts [84] has so far only been demonstrated in B6 mice (though not in the mHB).The consistency of the behavioral profiles between the present study and van der Goot et al. [19] give a first indication that subtypes exist in at least 129 mice but this requires further study. Thus, although further research is necessary to determine whether the identified subtypes reflect differential anxiety profiles, or represent differential responsivity to the test, the present results show that mice of various inbred strains may differ in their behavioral anxiety response.

This finding in itself is also relevant from another perspective, namely that defining animals on an individual level and incorporating this information in the analysis of results may contribute to the quality of animal experiments [19]. Lonsdorf and Merz [5] for example argued that subpopulations displaying contrasting response patterns may obscure the detection of significant differences at group level (i.e. a type II error).

Moreover, incorporating inter-individual variability in the design and analysis may enable researchers to better adhere to one of the fundamental principles of good experimental design of animal experiments: that all variables should be controlled, except that due to treatment [92]. Inter-individual variability – with its complex origin and elusive nature – has been proven a major factor undermining this principle [8]. This complexity makes it challenging to completely control for type of variability through increased standardization and therefore an increasing body of research has advocated the incorporation of this variability in the design and statistical analysis of animal experiments as an alternative way forward [93][94][95][96]. The presently applied approach as such may contribute to existing approaches advocating such incorporation.

With the benefits of this study stated, the findings also presented a number of limitations, which are discussed below. For one, the cluster stability of the presently identified clusters was low compared to the stability of the clusters identified by van der Goot et al. [19], despite the fact that the behavioral profiles largely overlapped between the two studies. This was unexpected, especially since we found highly stable clusters in a second (to be published) dataset obtained from testing the same strains in the same behavioral test.

The instability of the clusters at first glance contradicts the suggestion that the identified profiles reflect inter-individual variability in the

observed strains. An alternative explanation may however be found in the intricate properties of k-means cluster analysis.

In unsupervised cluster analyses, the stability of the clusters can be used to infer information about the reproducibility, or reliability of the clusters [87]. The bootstrapping procedure that was applied in the present study essentially compared cluster solutions of a large number of random subsets of the original data, with the rationale that clusters are stable if these subsamples produce similar results [69]. One of the characteristics of *k*-means cluster analysis, is that the starting point for the construction of the clusters is randomly selected every time the algorithm groups the data, and this starting point largely determines the partitioning of the remainder of the data into clusters [89]. In the current study, the clusters were close in size (53.9% of the mice in cluster A; 46.1% in cluster B). The near even distribution of mice between the clusters may therefore have caused the starting points to alternate between response type A and B at each of the bootstrap iterations, which in turn may have inadvertently accounted for the relatively instable bootstrap results.

Second, the present study also shows how experimenter effects can affect experimental results. The experimenter has been widely acknowledged as an uncontrollable factor in animal experimentation [20] [96][98][99][100]. Factors such as sex [101], familiarity with the experimenter [102] and experimenter experience [96] may all affect behavioral traits. Inter-observer variability constitutes another form of experimenter-induced variation [96][103]. In this study, the kappa statistic indicated moderate to good inter-observer reliability with on average a high percentage agreement (Section 2.4). Inter-observer reliability in itself however can again be affected by numerous factors such as experience, training, the rapidity of behavior, energy level of the observer and so on [96][104].

Automated tracking has been advocated as a means to overcome the incontrollable nature of this 'human element' and as such to increase the standardization of an experiment [103]. To our knowledge, fully automated scoring has unfortunately not yet been validated in the modified Hole Board, and doing so was beyond the scope of the present study.

Furthermore, each experimenter was allocated to one of the two animal rooms making it difficult to dissociate between experimenter-related and room effects. As described in Section 2.2, the humidity and temperature were comparable between the two animal rooms, but other factors could have played a role as well (i.e. barometric pressure, noise) [105]. A means to dissociate between these experimenter-related effects and room-effects in future research would be to have both experimenters observe data in both animal rooms.

Although the factor experimenter was controlled for in the residuals that were used for cluster analysis, the current study does not permit a definitive exclusion of the possibility that the found experimenter effects affected (some of the) variability that was found in the data. If anything, the experimenter effects in the present study emphasize the importance of accounting for this factor in experimental design, analysis and report of the results [97].

Third, the fact that the behavioral test was conducted in the housing room may have affected variability between individuals that were tested later in the experiment in comparison to animals that were assessed earlier on in the experiment. Research has demonstrated that ultra-sonic vocalizations may affect corticosterone responses and behavior in mice [106].

We have attempted to avoid a bias in our data of this potential confounding influence by randomizing test order across strains, within test day and batch. Furthermore, the random factor 'test order' did not contribute significantly to the variance in the models that assessed between strain differences, and test order was controlled for when obtaining the residuals that were used for clustering the data.

A final consideration with respect to the outcomes of this study is that the identified profiles only pertain to male mice, which limits the impact and conclusions of this study. As described in the section 'Animals and Housing' our rationale to including only males was driven by sample size requirements that were warranted by our utilized clustering approach and the associated heavy technical load of repeated testing of multiple inbred strains. Assessment of inter-individual variability in adaptive capacities of anxiety responses in both sexes however is essential, especially in the context of rodent models of anxiety. Anxiety disorders are more prevalent among women than in men [107] and the clinical course and treatment response are known to differ between sexes [108]. Mapping inter-individual variability in both sexes as such is pivotal for providing further insight in the underlying mechanisms that drive these differential responses and vulnerability [109].

5. Conclusion

This study empirically demonstrates, that inter-individual variability in habituation and sensitization of anxiety responses exists within males of three commonly used mouse inbred strains. The currently identified profiles are in line with previous findings, and as such suggest that they may be representative of subtypes of behavioral response in the observed strains. The three strains differ in innate emotionality. Whether the identified subgroups represent differential adaptive capacities regarding anxiety responses, or whether they represent individuals that are less responsive to the test may therefore differ between strains and requires further study. Also, further study is required to map inter-individual variability in female mice of these strains. The profiles identified in this study however provide a useful starting point for such further assessment.

6. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None

Acknowledgements

The skilled technical assistance Nicky van Kronenburg and Anja van der Sar is gratefully acknowledged.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.physbeh.2021.113503.

References

[1] H. Cohen, A.B. Geva, M.A. Matar, J. Zohar, Z. Kaplan, Post-traumatic stress behavioural responses in inbred mouse strains: can genetic predisposition explain phenotypic variability? Int. J. Neuropsychoph. 11 (2008) 331–349, https://doi.org/10.1017/S1461145707007912.

[2] J.M. Koolhaas, S.F. de Boer, C.M. Coppens, B. Buwalda, Neuroendocrinology of coping styles: towards understanding the biology of individual variation, Front. Neuroendocrinol. 31 (3) (2010) 307–321, https://doi.org/10.1016/j.yfrne.2010.04.001.

[3] A. Armario, R. Nadal, Individual differences and the characterization of animal models of psychopathology: a strong challenge and a good opportunity, Front. Pharmacol. 4 (2013) 137, https://doi.org/10.3389/fphar.2013.00137.

[4] I.R. Galatzer-Levy, G.A. Bonanno, D.E.A. Bush, J.E. LeDoux, Heterogeneity in threat extinction learning: substantive and methodological considerations for identifying individual differences in response to stress, Front. Behav. Neurosci. 7 (2013) 55, https://doi.org/10.3389/fnbeh.2013.00055, 10.3389/fnbeh.2013.00055.

[5] T.B. Lonsdorf, C.J. Merz, More than juist noise: inter-individual differences in fear acquisition, extinction and fear in humans – biological, experiential, temperamental factors, and methodological pitfalls, Neurosci. Biobehav. Rev. 80 (2017) 703–728, https://doi.org/10.1016/j.neurobiorev.2017.07.007.

[6] H. Einat, I. Ezer, N. Kara, C. Belzung, Individual responses of rodents in modelling of affective disorders and in their treatment: prospective review, Acta. Neuropsychiatr. 30 (6) (2018) 323–333, https://doi.org/10.1017/neu.2018.4.

[7] R. Lathe, The individuality of mice, Gene. Brain Behav. 3 (2004) 317–327, https://doi.org/10.1111/j.1601-183X.2004.00083.x.

[8] B. Voelkl, N.S. Altman, A. Forsman, W. Forstmeier, J. Gurevitch, I. Jaric, N. A. Karp, M.J. Kas, H. Schielzeth, T. Van de Casteele, H. Würbel, Reproducibility of animal research in the light of biological variation, Nat. Rev. Neurosci. 2020; 21 (2020) 384–393, https://doi.org/10.1038/s41583-020-0313-3.
[9] L. Kazavchinsky, A. Dafna, H. Einat, Individual variability in female and male mice in a test-retest protocol of the forced swim test, J. Pharmacol. Toxicol. Meth. 95 (2019) 12–15, http://doi.org/j.vascn.2018.11.007.
[10] E.M. Eisenstein, D. Eisenstein, A behavioral homeostasis theory of habituation and sensitization: II. Further developments and predictions, Rev. Neurosci. 17 (5) (2006) 533–557, https://doi.org/10.1515/REVNEURO.2006.17.5.533.
[11] A.R. Salomons, J.A.K.R. van Luijk, N.R. Reinders, S. Kirchhoff, S.S. Arndt, F. Ohl, Identifying emotional adaptation: behavioural habituation to novelty and immediate early gene expression in two inbred mouse strains, Gene. Brain Behav. 9 (1) (2010) 1–10, https://doi.org/10.1111/j.1601-183X.2009.00527.x.
[13] H. Boleij, A.R. Salomons, M. van Sprundel, S.S. Arndt, F. Ohl, Not all mice are equal: welfare implications of behavioural habituation profiles in four 129 mouse substrains, PLoS ONE 7 (8) (2012) e42544, https://doi.org/10.1371/journal.pone.0042544.
[14] A.R. Salomons, G. Bronkers, S. Kirchhoff, S.S. Arndt, F. Ohl, Behavioural habituation to novelty and brain area specific immediate early gene expression in female mice of two inbred strains, Behav. Brain Res. 215 (1) (2010) 95–101, https://doi.org/10.1016/j.bbr.2010.06.035.
[15] A.R. Salomons, T. Kortleve, N.R. Reinders, S. Kirchhoff, S.S. Arndt, F. Ohl, Susceptibility of a potential animal model for pathological anxiety to chronic mild stress, Behav. Brain Res. 209 (2) (2010) 241–248, https://doi.org/10.1016/j.bbr.2010.01.050.
[16] A.R. Salomons, S.S. Arndt, M. Lavrijsen, F Kirchhoff, Ohl, Expression of crfr1 and glu5r mrna in different brain areas following repeated testing in mice that differ in habituation behavior, Behav. Brain Res. 246 (2013) 1–9, https://doi.org/10.1016/j.bbr.2013.02.023.
[17] G.W.M. Bothe, V.J. Bolivar, M.J. Vedder, J.G. Geistfeld, Behavioral differences among fourteen inbred mouse strains commonly used as disease models, Comp. Med. 55 (4) (2005) 326–334. PMID: 16158908.
[18] C. Belzung, G. Griebel, Measuring normal and pathological anxiety-like behavior in mice: a review, Behav. Brain Res. 125 (1–2) (2001) 141–149, https://doi.org/10.1016/S0166-4328(01)00291-1.
[19] M.H. Van der Goot, H. Boleij, J. van den Broek, A.R. Salomons, S.S. Arndt, H. A. van Lith, An individual based, multidimensional approach to identify emotional reactivity profiles in inbred mice, J. Neurosci. Meth. (2020), https://doi.org/10.1016/j.neumeth.2020.108810.
[20] J.C. Crabbe, D. Wahlsten, B.C. Dudek, Genetics of mouse behavior: interactions with laboratory environment, Sci. 284 (5420) (1999) 1670–1672, https://doi.org/10.1126/science.284.5420.1670.
[21] J.P. Garner, Stereotypies and other abnormal repetitive behaviors: potential impact on validity, reliability, and replicability of scientific outcomes, ILAR. J. 46 (2) (2005) 106–117, https://doi.org/10.1093/ilar.46.2.106.
[22] W.Y. Tam, K.-.K. Cheung, Phenotypic characteristics of commonly used mouse inbred strains, J. Mol. Med. (2020), https://doi.org/10.1007/s00109-020-1953-4. Jul 25.
[23] D.C. Rogers, D.N.C. Jones, P.R. Nelson, C.M. Jones, Ch.A. Quilter, T.L. Robinson, J.J. Hagan, Use of shirpa and discriminant analysis to characterize marked differences in the behavioural phenotype of six inbred mouse strains, Behav. Brain Res. 105 (2) (1999) 207–217, https://doi.org/10.1016/s0166-4328(99)00072-8.
[24] V.J. Bolivar, B.J. Caldarone, A.A. Reilly, L. Flaherty, Habituation of activity in an open field: a survey of inbred strains and f1 hybrids, Behav. Genet. 30 (2000) 285–293, https://doi.org/10.1023/A:1026545316455.
[25] G. Griebel, C. Belzung, G. Perrault, D.J. Sanger, Differences in anxiety related behaviours and in sensitivity to diazepam in inbred and outbred strains of mice, Psychopharmacol. (Berl.) 148 (2) (2000) 164–170, https://doi.org/10.1007/s002130050038.
[26] L. de Visser, R. van den Bos, W.W. Kuurman, M.J.H. Kas, B.M. Spruijt, Novel approach to the behavioural characterization of inbred mice: automated home cage observations, Gene. Brain Behav. 5 (6) (2006) 458–466, https://doi.org/10.1111/j.1601-183X.2005.00181.x.
[27] F. Ohl, Testing for anxiety, Clin. Neurosci. Res. 3 (4–5) (2003) 233–238, https://doi.org/10.1016/S1566-2772(03)00084-7.
[28] S.M. Korte, Corticosteroids in relation to fear, anxiety, and psychopathology, Neurosci. Biobehav. Rev. 25 (2) (2001) 117–142, https://doi.org/10.1016/s0149-7634(01)00002-1.
[29] A. Sgoifo, S.F. de Boer, J. Haller, J.M. Koolhaas, Individual differences in plasma catecholamine and corticosterone stress responses of wild-type rats, Physiol. Behav. 60 (6) (1996) 1403–1407, https://doi.org/10.1016/S0031-9384(96)00229-6.
[30] F. Rougé-Pont, V. Deroche, M. Le Moal, P.V. Piazza, Individual differences in stress-induced dopamine release in the nucleus accumbens are influenced by corticosterone, Eur. J. Neurosci. 10 (12) (1998) 3903–3907, https://doi.org/10.1046/j.1460-9568.1998.00438.x.
[31] J.F. Cockrem, Individual variation in glucocorticoid stress responses in animals, Gen. Comp. Endocrinol. 181 (2013) 45–58, https://doi.org/10.1016/j.ygcen.2012.11.025.
[32] M. Jakovcevski, M. Schachner, F. Morellini, Individual variability in the stress response of c57bl/6 j male mice correlates with trait anxiety, Gene. Brain Behav. 7 (2008) 235–243, https://doi.org/10.1111/j.1601-183X.2007.00345.x.
[33] M. Jakovcevski, M. Schachner, F. Morellini, Susceptibility to the long-term anxiogenic effects of an acute stressor is mediated by the activation of the glucocorticoid receptors, Neuropharmacol. 61 (8) (2011) 1297–1305, https://doi.org/10.1016/j.neuropharm.2011.07.034.
[34] C. Nasca, B. Bigio, D. Zelli, F. Nicoletti, B.S. McEwen, Mind the gap: glucocorticoids modulate hippocampal glutamate tone underlying individual differences in stress susceptibility, Mol. Psychiatr. 20 (2015) 755–763, https://doi.org/10.1038/mp.2014.96.
[35] N. Percie du Sert, V. Hurst, A. Ahluwalia, S. Alam, M.T. Avey, M. Baker, W. J. Browne, A. Clark, I.C. Cuthill, U. Dirnagl, M. Emerson, P. Garner, S.T. Holgate, D.W. Howells, N.A. Karp, K. Lidster, C.J. MacCallum, M. Macleod, O. Petersen, F. Rawle, P. Reynolds, K. Rooney, E.S. Sena, S.D. Silberberg, T. Steckler, H. Würbel, The arrive guidelines 2.0: updated guidelines for reporting animal research, BMC. Vet. Res. 16 (2020) 242, https://doi.org/10.1186/s12917-020-02451-y.
[36] N. Percie du Sert, A. Ahluwalia, S. Alam, M.T. Avey, M. Baker, W.J. Browne, A. Clark, I.C. Cuthill, U. Dirnagl, M. Emerson, P. Garner, S.T. Holgate, D. W. Howells, V. Hurst, N.A. Karp, S.E. Lazic, K. Lidster, C.J. MacCallum, M. Macleod, E.J. Pearl, O.H. Petersen, F. Rawle, P. Reynolds, K. Rooney, E. S. Sena, S.D. Silberberg, T. Steckler, H. Würbel, Reporting animal research: explanation and elaboration for the arrive guidelines 2.0, PLos Biol. 18 (7) (2020), e3000411, https://doi.org/10.1371/journal.pbio.3000411.
[37] S. Dolnicar, B. Grün, F. Leisch, Increasing sample size compensates for data problems in segmentation studies, J. Bus. Res. 69 (2) (2016) 992–999, https://doi.org/10.1016/j.jbusres.2015.09.004.
[38] S.S. Arndt, M.C. Laarakker, H.A. van Lith, F.J. van der Staay, E. Gieling, A. R. Salomons, J. van 't Klooster, F. Ohl, Individual housing of mice – impact on behavior and stress responses, Phys. Behav. 97 (3) (2009) 385–393, https://doi.org/10.1016/j.physbeh.2009.03.0008.
[39] M.F.W. Festing, V. Baumans, R.D. Combes, M. Halder, C.F.M. Hendriksen, B. R. Howard, D.P. Lovell, G.J. Moore, P. Overend, M.S. Wilson, Reducing the of laboratory animals in biomedical research: problems and possible solutions, Altern. Lab. Anim. 26 (3) (1998) 283–301. PMID: 26042346.
[40] R. Shaw, M.F.W. Festing, I. Peers, L. Furlong, Use of factorial designs to optimize animal experiments and reduce animal use, ILAR. J. 43 (4) (2002) 223–232, https://doi.org/10.1093/ilar.43.4.223.
[41] T. Buch, K. Moos, F.M. Ferreira, H. Fröhlich, C. Gebhard, A. Tresch, Benefits of a factorial design focussing on inclusion of female and male animals in one experiment, J. Mol. Med. 97 (2019) 871–877, https://doi.org/10.1007/s00109-019-01774-0.
[43] K. Gouveia, J. Hurst, Reducing mouse anxiety during handling: effect of experience with handling tunnels, PLoS ONE 8 (6) (2013) e66401, https://doi.org/journal.pone.0066401.
[44] F. Ohl, I. Sillaber, E. Binder, M.E. Keck, F. Holsboer, Differential analysis of behavior and diazepam-induced alterations in c57bl/6 n and balb/c mice using the modified hole board test, J. Psychiatr. Res. 35 (3) (2001) 147–154, https://doi.org/10.1016/S0022-3956(01)00017-6.
[45] M. Labots, H.A. van Lith, F. Ohl, S.S. Arndt, The modified hole board –measuring behavior, cognition and social interaction in mice and rats, J. Vis. Exp. 98 (2015) e52529, https://doi.org/10.3791/52529.
[46] D.V. Cicchetti, The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements, J. Clin. Exp. Neuropsyc. 23 (5) (2001) 695–700, https://doi.org/10.1076/jcen.23.5.695.1249.
[47] S. Honma, K.I. Honma, T. Shirakawa, T. Hiroshige, Rhythms in behaviors, body temperature and plasma corticosterone in scn lesioned rats given methamphetamine, Physiol. Behav. 44 (2) (1988) 247–255, https://doi.org/10.1016/0031-9384(88)90146-1.
[49] M. Dürschlag, H. Würbel, M. Stauffacher, D. Von Holst, Repeated blood collection in the laboratory mouse by tail incision – modification of an old technique, Phsyiol. Behav. 60 (6) (1996) 1565–1568, https://doi.org/10.1016/S0031-9384(96)00307-1.
[50] M.C. Laarakker, F. Ohl, H.A. van Lith, 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. 38 (2) (2008) 159–184, https://doi.org/10.1007/s10519-007-9188-6.
[51] M.C. Laarakker, H.A. van Lith, F. Ohl, Behavioral characterization of A/J and C57BL/6 J mice using a multidimensional test: association between bloodplasma and brain magnesium-ion concentration with anxiety, Physiol. Behav. 102 (2) (2011) 205–219, https://doi.org/10.1016/j.physbeh.2010.10.019.
[52] M. Labots, M.C. Laarakker, D. Schetters, S.S. Arndt, H.A. van Lith, An improved procedure for integrated behavioral z-scoring illustrated with modified hole board behavior of male inbred laboratory mice, J. Neurosci. Meth. 293 (2018) 375–388, https://doi.org/10.1016/j.jneumeth.2017.09.003.
[53] J. Guilloux, M. Seney, N. Edgar, E. Sibille, Integrated behavioral z-scoring increases the sensitivity and reliability of behavioral phenotyping in mice: relevance to emotionality and sex, J. Neurosci. Meth. 197 (1) (2011) 21–31, https://doi.org/10.1016/j.jneumeth.2011.01.019.
[54] R.D. Cook, S. Weisberg, Residuals and Influence in Regression, Chapman and Hall, New York, 1982. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/37076.
[55] R Core Team, R: A language Environment For Statistical Computing, R foundation for Statistical Computing, Vienna, Austria, 2020. https://www.R-project.org/.
[56] Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., R Core Team, 2020. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1.-147, URL: https://CRAN.R-project.org/package=nlme.

[57] M.E. Brooks, K. Kristensen, K.J. van Benthem, A. Magnusson, C.W. Berg, A. Nielsen, H.J. Skaug, M. Maechler, B.M. Bolker, glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modelling, R. J. 9 (2) (2017) 378–400.

[58] R.R. Sokal, F.J. Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research, W.H. Freeman and Co., New York, NY, 1995. Third edition.

[59] A.F. Zuur, E.N. Ieno, N.J. Walker, A.A. Saveliev, G.M. Smith, Mixed Effects Models and Extensions in Ecology with R, Springer, New York, NY, 2009, https://doi.org/10.1007/978-0-387-87458-6.

[60] R. Lenth, Emmeans: Estimated Marginal Means, Aka Least-Squares Means, 2020. R package version 1.4.7, https://CRAN.R-project.org/package-emmeans.

[61] Z. Šidák, Rectangular confidence regions for the means of multivariate normal distributions, J. Am. Stat. Assoc. 62 (318) (1967) 626–633, https://doi.org/10.1080/01621459.1967.1048295.

[62] D. Wahlsten, Sample size. Mouse Behavioral testing: How to Use Mice in Behavioral neuroscience, 1st Edn., Academic Press, Elsevier Inc., London, U.K., 2011, pp. 75–105.

[63] S.T. Bate, R.A. Clark, The Design and Statistical Analysis of Animal Experiments, Cambridge University Press, UK, 2014.

[64] C. Genolini, X. Alacoque, M. Sentenac, C. Arnaud, Kml and kml3d: r-packages to cluster longitudinal data, J. Stat. Softw. 65 (4) (2015) 1–34. http://www.jstatsoft.org/v65/i04/.

[65] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2019. Cluster: cluster analysis basics and extensions. R package version 2.1.0.

[66] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Statist. Soc. B. 63 (2) (2002) 411–423, https://doi.org/10.1111/1467-9868.00293.

[67] Kryszczuk, K., Hurley, P., 2010. Estimation of the number of clusters using multiple clustering validity indices. In: El Gayar, N., Kittler, J., Roli, F. (eds). Multiple Classifier Systems. MCS 2010. Lecture Notes in Computer Science, vol 5997. Springer, Berlin, Heidelberg. 10.1007.

[68] S. Wahl, S. Krug, C. Then, et al., Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the fto obesity risk allele, Metabolomics 10 (3) (2014) 386–401, https://doi.org/10.1007/s11306-013-0586-x.

[69] J. Clatworthy, D. Buick, M. Hankins, J. Weinman, R. Home, The use and reporting of cluster analysis in health psychology: a review, Br. J. Heal. Psychol. 10 (2005) 329–358, https://doi.org/10.1348/135910705X25697. Pt 310.1348/135910705X25697.

[70] P. Ardayfio, K. Kim, Anxiogenic-like effect of chronic corticosterone in the light-dark emergence task in mice, Behav. Neurosci. 120 (2) (2006) 249–256, https://doi.org/10.1037/0735-7044.120.2.249.

[71] T.P. O'Leary, R.K. Gunn, R.E. Brown, What are we measuring when we test strain differences in anxiety in mice? Behav. Genet. 43 (1) (2013) 34–50, https://doi.org/10.1007/s10519-012-9572-8.

[72] K. Ebner, N. Singewald, Individual differences in stress susceptibility and stress inhibitory mechanisms, Curr. Opin. Behav. Sci. 14 (2017), https://doi.org/10.1016/j.cobeha.2016.11.016, 65-64.

[73] M. Weger, C. Sandi, High anxiety trait: a vulnerable stress phenotype for stress-induced depression, Neurosci. Biobehav. Rev. 87 (2018) 27–37, https://doi.org/10.1016/j.neubiorev.2018.01.012.

[74] P.V. Piazza, S. Maccari, J.M. Deminiere, M. LeMoal, P. Mormede, H. Simon, Corticosterone levels determine individual vulnerability to amphetamine self-administration, Proc. Natl. Acad. Sci. U S A 88 (6) (1991) 2088–2092, https://doi.org/10.1073/pnas.88.6.2088.

[75] M. Loos, B. Koopmans, E. Aarts, G. Maroteaux, S. van der Sluis, Neuro-BSIK Mouse Phenomics Consortium, M. Verhage, A.B. Smit, Within-strain variation in behavior differs consistently between common inbred strains of mice, Mamm. Genome. 26 (7–8) (2015) 348–354, https://doi.org/10.1007/s00335-015-9578-7.

[76] C. Ducottet, C. Belzung, Behaviour in the elevated-plus-maze predicts coping after subchronic mild stress in mice, Physiol. Behav. 81 (3) (2004) 417–426, https://doi.org/10.1016/j.physbeh.2004.01.013.

[77] M.N. Cook, V.J. Bolivar, Behavioral differences among 129 substrains: implication for knockout and transgenic mice, Behav. Neurosci. 116 (4) (2002) 600–611, https://doi.org/10.1037/0735-7044.116.4.600.

[78] M. Pratte, M. Jamon, Detection of social approach in inbred mice, Behav. Brain Res. 203 (2009) 54–64, https://doi.org/10.1016/j.bbr.2009.04.011.

[79] R. Kakihana, J.A. Moore, Circadian rhythm of corticosterone in mice: the effect of chronic consumption of alcohol, Phsychopharmacol. 46 (1976) 301–305. https://10.1007/BF00421118.

[80] H. Neuman, J.W. Debelius, R. Knight, O. Koren, Microbial endocrinology: the interplay between the microbiota and the endocrine system, FEMS. Microbiol. Rev. 39 (4) (2015) 509–521, https://doi.org/10.1093/femsre/fuu010.

[81] T.L. Jensen, M.K. Kiersgaard, D.B. Sorensen, L.F. Mikkelsen, Fasting of mice: a review, Lab. Anim. 47 (4) (2013) 225–240, https://doi.org/10.1177/0023677213501659.

[82] V. Brinks., M. van der Mark, R. de Kloet, M. Oitzl, Emotion and cognition in high and low stress sensitive mouse strains: a combined neuroendocrine and behavioral study in balb/c and c57bl/6 mice, Front. Behav. Neurosci. 1 (8) (2007), https://doi.org/10.3389/neuro.08.008.2007.

[83] S.B. Sartori, R. Landgraf, N. Singewald, The clinical implications of mouse models of enhanced anxiety, Fut. Neurol. 6 (4) (2011) 531–571, https://doi.org/10.2217/fnl.11.34.

[84] L. Lewejohann, B. Zipser, N. Sachser, „Personality" in laboratory mice used for biomedical research: a way of understanding variability? Dev. Psychobiol. 53 (6) (2011) 624–630, https://doi.org/10.1002/dev.20553.

[85] Keshavarz, M., Krebs-Watson, R., Refki, P., Savriama, Y., Zhang, Y., Guenther, A., Brückl, T.M., Binder, E.B., Tautz, D., 2020. Natural copy number variation differences of tandemly repeated small nucleolar rnas in the prader-willi syndrome genomic region regulate individual behavioral responses in mammals. bioRxiv 476010, 10.1101/476010.

[86] J. Freund, A.M. Brandmaier, L. Lewejohann, I. Kirste, M. Kritzler, A. Krüger, N. Sachser, U. Lindenberger, G. Kempermann, Emergence of individuality in genetically identical mice, Sci. 340 (6133) (2013) 756–759, https://doi.org/10.1126/science.1235294.

[87] M. Labots., X. Zheng, G. Moattari, F. Ohl, H.A. van Lith, Effects of light regime and substrain on behavioral profiles of male c57bl/6 mice in three tests of unconditioned anxiety, J. Neurogenet. 30 (4) (2016) 306–315, https://doi.org/10.1080/01677063.2016.1249868.

[89] V.Y. Kiselev, T.S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell rna-seq data, Nat. Rev. Genet. 20 (2019) 273–282, https://doi.org/10.1038/s41576-018-0088-9.

[92] M.F.W. Festing, Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research, ILAR. J. 55 (3) (2014) 399–404, https://doi.org/10.1093/ilar/ilu036.

[93] S.H. Richter, J.P. Garner, C. Auer, J. Kunert, H. Wuerbel, Systematic variation improves reproducibility of animal experiments, Nat. Meth. 7 (2010) 167–168, https://doi.org/10.1038/nmeth0310-167.

[94] S.H. Richter, Systematic heterogenization for better reproducibility in animal experimentation, Lab. Anim. 46 (2017) 343–349, https://doi.org/10.1038/laban.1330.

[95] N. Kafkafi, J. Agassi, E.J. Chesler, Reproducibility and replicability of rodent phenotyping in preclinical studies, Neurosci. Biobehav. Rev. 87 (2018) 218–232, https://doi.org/10.1016/j.neubiorev.2018.01.003.

[96] N.M. Bello, D.G. Renter, Reproducible research from noisy data: revisiting key statistical principles for the animal sciences, J. Dairy Sci. 101 (2018) 5679–5701, https://doi.org/10.3168/jds.2017-13978.

[97] M. Bohlen, E.R. Hayes, B. Bohlen, B.D. Bailoo, J.C. Crabbe, D. Wahlsten, Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol, Behav. Brain Res. 217 (2014) 46–54, https://doi.org/10.1016/j.bbr.2014.06.017.

[98] D. Wahlsten, P. Metten, T.J. Philips, S.L. Boehm, S. Burkhart-Kasch, J. Dorow, Different data from different labs: lessons from studies of gene-environment interaction, J. Neurobiol. 54 (2003) 283–311, https://doi.org/10.1002/neu.10173.

[99] E.J. Chesler, S.G. Wilson, W.R. Lariviere, S.L. Rodriguez-Zas, J.S. Mogil, Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive, Neurosci. Biobehav. Rev. 26 (2002) 907–923, https://doi.org/10.1016/s0149-7634(02)00103-3.

[100] E.J. Chesler, S.G. Wilson, W.R. Lariviere, S.L. Rodriguez-Zas, J.S. Mogil, Influences of laboratory environment on behavior, Nat. Neurosci. 5 (2002) 1101–1102, https://doi.org/10.1038/nn1102-1101.

[101] R.E. Sorge, L.J. Martin, K.A. Isbester, S.G. Sotocinal, S. Rosen, A.H. Tuttle, J. S. Wieskopf, E.L. Acland, A. Dokova, B. Kadoura, P. Leger, J.C.S. Mapplebeck, M. McPhail, A. Delaney, G. Wigerblad, A.P. Schumann, T. Quinn, J. Frasnelli, C. I. Svensson, W.F. Sternberg, J.S. Mogil, Olfactory exposure to males, including men, causes stress and related analgesia in rodents, Nat. Meth. 11 (2014) 629–632, https://doi.org/10.1038/nmeth.2935.

[102] K.S. Van Driel, J.C. Talling, Familiarity increases consistency in animal tests, Behav. Brain Res. 159 (2005) 243–245, https://doi.org/10.1016/j.bbr.2004.11.005.

[103] S.H. Richter, Automated home-cage testing as a tool to improve reproducibility of behavioral research? Front. Neurosci. 14 (2020) 383, https://doi.org/10.3389/fnins.2020.00383.

[104] A.B. Kaufman, R. Rosenthal, Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour, Anim. Behav. 78 (6) (2009) 1478–1491, https://doi.org/j.anbehav.2009.09.014.

[105] J.S. Mogil, Laboratory environment factors and pain behavior: the relevance of unknown unknowns to reproducibility and translation, Lab Anim. (N.Y.) 46 (2017) 136–141, https://doi.org/10.1038/laban.1223.

[106] A.C. Niemczura, J.M. Grimsley, C. Kim, A. Alkhawaga, A. Poth, A. Carvalho, J. J. Wenstrup, Physiological and behavioral response to vocalization playback in mice, Front. Behav. Neurosci. 14 (2020) 155, https://doi.org/10.3389/fnbeh.2020.00155.

[107] R. Zender, E. Olshansky, Women's mental health: depression and anxiety, Nurs. Clin. North. Am. 44 (3) (2009) 335–364, https://doi.org/10.1016/j.cnur.2009.06.002.

[108] N.C. Donner, C.A. Lowry, Sex differences in anxiety and emotional behavior, Eur. J. Physiol. 465 (2013) 601–626, https://doi.org/10.1007/s00424-013-1271-7.

[109] P.M. Pitychoutis, E.G. Pallis, H.G. Mikail, Z. Papadopoulou-Daifoti, Individual differences in novelty-seeking predict differential responses to chronic antidepressant treatment through sex- and phenotype-dependent neurochemical signatures, Behav. Brain Res. 223 (1) (2011) 154–168, https://doi.org/10.1016/j.bbr.2011.04.036.

Table S1. Mean body weight at arrival (grams ± SD) for each batch [1-5] of each strain.

Batch	Strain	Date of arrival	Body weight at arrival (mean gr ± SD)	Range (min – max)
1	C	24-5-2018	18.82 ± 0.55	16.5 – 21.2
1	129S2	24-5- 2018	22.80 ± 0.72	18.2 – 26.7
2	C	31-5-2018	21.45 ± 0.33	20.1 – 23.3
2	129S2	31-5-2018	23.82 ± 0.96	20 – 29.1
2	B6N	31-5-2018	22.74 ± 0.58	19.6 – 24.7
3	C	6-6-2018	17.71 ± 0.30	16.4 – 19
3	129S2	6-6-2018	22.16 ± 0.48	19.7 – 24.3
3	B6N	6-6-2018	19.09 ± 0.27	17.8 – 20.3
4	C	13-6-2018	17.99 ± 0.39	16.2 – 19.5
4	129S2	13-6-2018	27.4 ± 0.33	26.3 – 29.7
4	B6N	13-6-2018	21.72 ± 0.54	20.1 – 24.5
5	B6N	4-7-2018	20.56 ± 0.44	18.2 – 22.6

Table S2. Behavioral variables measured in the mHB and used for composition of z-scores in this publication.

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
- *Risk assessment*	Total number of stretched attends	z
	Latency until first stretched attend	-z
- *Arousal*	Total number of self-groomings	z
	Latency until the first self-grooming	-z
	Percentage of time self-grooming	z
	Total number of boli	z
	Latency until first boli is produced	-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 S3. Strain differences: Effects of different explanatory variables for each behavioral dimension/pCORT. Behavioral dimensions were analyzed with generalized linear mixed models (GLMMs) using a 3 (strain) x 2 (experimenter) x 4 (trials) mixed factorial design. Strain, experimenter, trial and their interactions were included as fixed predictors. Mouse identity (ID), mouse slope (trial nested in ID), batch and testorder were included as random factors.
pCORT: linear mixed model (LMM) using a 3 (strain) x 4 (technician) x 3 (sampling moment) mixed factorial design. Strain, sampling moment and bio-technician were included as fixed predictors, including their interactions. Day of test was included as fixed covariate. Mouse identity, mouse slope (sampling moment nested in ID), batch, testorder were included as random factors. Significant effects are highlighted in bold.

Dimension	Explanatory variables	F	df	P
(a)Avoidance	Strain S	17.44	2, 111	**< 0.0001**
	Trial T	11.04	3, 329	**< 0.0001**
	Experimenter E	8.61	1, 111	**0.0041**
	S*T	8.51	6, 329	**<0.0001**
	S*E	1.14	2, 111	0.3236
	T*E	0.70	3, 329	0.5529
	S*T*E	2.12	6, 329	0.0507
	Mouse identity[r]	117.68 [b]	14	**< 0.0001**
	Mouse slope[r]	21.01 [b]	16	**< 0.0001**
	batch[r]	0.25 [b]	19	0.9686
	testorder[r]	0.00 [b]	22	1.000
(b)Risk assessment				
glmmTMB	Strain S[a]	186.53[b]	2	**<.0001**
	Trial T[a]	678.71 [b]	3	**<.0001**
	Experimenter E[a]	3.59 [b]	1	0.0580
	S*T[a]	175.72 [b]	6	**<.0001**
	S*E[a]	4.56 [b]	2	0.1022
	T*E[a]	23.20 [b]	3	**0.0001**
	S*T*E[a]	12.03 [b]	6	0.0613
	Mouse identity[r]	14.95 [b]	15	**0.0001**
	Mouse slope[r]	1.58 [b]	17	0.4535
	batch[r]	-*	-	-
	testorder[r]	-	-	-
(c)Arousal	Strain S	3.49	2, 111	**0.0339**
	Trial T	20.66	3, 329	**<.0001**
	Experimenter E	9.15	1, 111	**0.0031**
	S*T	3.57	6, 329	**0.0019**
	S*E	0.75	2, 111	0.4732
	T*E	2.45	3, 329	0.0631
	S*T*E	1.79	6, 329	0.1011
	Mouse identity[r]	22.82	14	**<0.0001**
	Mouse slope[r]	13.08	16	**0.0014**

	batch[r]	0.87	19	0.8328
	testorder[r]	0.01	22	0.9998
(d)Exploration	Strain S	38.15	2, 111	**< 0.0001**
	Trial T	79.00	3, 329	**< 0.0001**
	Experimenter E	114.95	1, 111	**< 0.0001**
	S*T	3.73	6, 329	**0.0113**
	S*E	0.34	2, 111	0.7117
	T*E	5.08	3, 329	**0.0019**
	S*T*E	0.84	6, 329	0.5367
	Mouse identity[r]	94.83 [b]	14	**< 0.0001**
	Mouse slope[r]	39.59 [b]	16	**< 0.0001**
	batch[r]	0.38 [b]	19	0.9445
	testorder[r]	0.05 [b]	22	0.9968
(e)Locomotion	Strain S	183.02	2, 111	**< 0.0001**
rank transformed	Trial T	18.91	3, 329	**< 0.0001**
	Experimenter E	5.26	1, 111	**0.0237**
	S*T	20.94	6, 329	**< 0.0001**
	S*E	3.50	2, 111	**0.0336**
	T*E	2.36	3, 329	0.0711
	S*T*E	0.76	6, 329	0.5977
	Mouse identity[r]	111.56[b]	14	**< 0.0001**
	Mouse slope[r]	27.61 [b]	16	**< 0.0001**
	batch[r]	1.41 [b]	19	0.7026
	testorder[r]	na	Na	Na
(f) pCORT	Strain S	21.81	2, 321	**< 0.0001**
	Sampling time (T)	263.98	2, 321	**< 0.0001**
	Biotechnician (B)	0.31	3, 321	0.8173
	S*T	4.67	4, 321	**0.0011**
	S*B	Na	Na	na
	T*B	Na	Na	na
	S*T*B	Na	Na	na
	Day of test (D)	0.12	1, 321	0.7231
	Mouse identity[r]	1.87	15	0.1718
	Mouse slope[r]	0.00	17	1.000
	batch[r]	Na	Na	na
	testorder[r]	Na	Na	na

[r]Mouse identity, mouse slope, batch and testorder were used as random factors; the statistical significance of these factors was calculated by likelihood-ratio tests and thus [b]Chi-square values are reported. [a]Analyses conducted with glmmTMB: [b]main and interaction effects reported with Chi Square values. Na = model did not converge.

Table S4. *Post hoc* within strain comparisons of the estimated marginal means between trials 1 and 4 for each behavioral dimension, and between sampling moments for pCORT. Behavioral dimensions: adjusted $\alpha = 0.016952$ for comparison between trial 1 and 4. pCORT): adjusted $\alpha = 0.012741$ for comparison between sampling moments. Significant contrasts are highlighted in bold.

Dimension		Estimate ± SEM	$t_{(df)}$	*P*	Cohens *d*	Lower and upper limits of 95%CI
(a)Avoidance						
Trial 1 vs 4	C	0.905 ± 0.137	$6.625_{(111)}$	**< 0.0001**	3.312	2.228, 4.396
	B6N	0.224 ± 0.135	$1.660_{(111)}$	0.0998	0.819	-0.165, 1.803
	129S2	-0.457 ± 0.140	$-3.249_{(111)}$	**0.0015**	-1.670	-2.712, -0.628
(b)Risk assessment						
Trial 1 vs 4	C	2.195 ± 0.111	$19.760_{(432)}$	**< 0.0001**	5.895	5.189, 6.602
	B6N	0.789 ± 0.083	$8.939_{(432)}$	**< 0.0001**	2.120	1.633, 2.608
	129S2	1.809 ± 0.114	$15.802_{(432)}$	**< 0.0001**	4.860	4.174, 5.546
(c)Arousal						
Trial 1 vs 4	C	-0.343 ± 0.107	$-3.195_{(329)}$	**0.0015**	-0.784	-1.272, -0.298
	B6N	-0.388 ± 0.112	$-3.462_{(329)}$	**0.0006**	-0.888	-1.397, -0.379
	129S2	-0.723 ± 0.111	$-6.528_{(329)}$	**< 0.0001**	-1.654	-2.169, -1.140
(d)Exploration						
Trial 1 vs 4	C	-0.780 ± 0.067	$-11.725_{(329)}$	**< 0.0001**	-2.591	-3.070, -2.113
	B6N	-0.512 ± 0.075	$-6.830_{(329)}$	**< 0.0001**	-1.701	-2.208, -1.194
	129S2	-0.538 ± 0.096	$-5.577_{(329)}$	**< 0.0001**	-1.787	-2.432, -1.142
(e)Locomotion *rank transformed*						
Trial 1 vs 4	C	-84.88 ± 19.8^{r}	$-4.295_{(329)}$	**< 0.0001**	-0.971	-1.421, -0.520
	B6N	149.95 ± 13.9^{r}	$1.780_{(329)}$	**< 0.0001**	1.715	1.375, 2.054
	129S2	-11.17 ± 20.1^{r}	$-0.557_{(329)}$	0.5780	-0.128	-0.579, 0.324
(f)pCORT						
Time 1 vs 2	C	-1.10 ± 0.31	$-3.586_{(40)}$	**0.0009**	-1.389	-2.194, -0.585
	B6N	-1.32 ± 0.33	$-4.065_{(39)}$	**0.0002**	-1.676	-2.541, -0.812
	129S2	-1.95 ± 0.47	$-4.170_{(38)}$	**0.0002**	-2.460	-3.701, -1.219
Time 1 vs 3	C	-0.25 ± 0.09	$-2.808_{(40)}$	**0.0077**	-0.320	-0.555, -0.086
	B6N	-0.05 ± 0.05	$-1.012_{(39)}$	0.3182	-0.060	-0.182, 0.061
	129S2	-0.14 ± 0.08	$-1.748_{(38)}$	0.0889	-0.187	-0.406, 0.031
Time 2 vs 3	C	0.85 ± 0.27	$3.184_{(40)}$	**0.0028**	1.069	0.375, 1.762
	B6N	1.28 ± 0.32	$4.019_{(39)}$	**0.0003**	1.616	0.772, 2.460
	129S2	1.80 ± 0.44	$4.102_{(38)}$	**0.0002**	2.273	1.110, 3.435

Table S5. *Post hoc* between strain comparisons of the estimated marginal means on each trial (behavioral dimensions) or on each sampling moment (pCORT). Behavioral dimensions: adjusted α = 0.016952 for trials 1 and 4, adjusted α = 0.025321 for trials 2 and 3. pCORT: adjusted α = 0.012741. Significant contrasts are highlighted in bold.

Dimension		Estimate ± SEM	$t_{(df)}$	P	Cohen's *d*	Lower and upper limits of 95%CI
(a)Avoidance						
Trial 1	C vs B6N	0.648 ± 0.178	$3.646_{(111)}$	**0.0004***	2.372	1.045, 3.699
	C vs 129S2	1.529 ± 0.162	$9.422_{(111)}$	**< 0.0001*****	5.595	4.203, 6.987

	B6N vs 129S2	0.881 ± 0.152	$5.815_{(111)}$	**< 0.0001*****	3.223	2.044, 4.402
Trial 2	C vs B6N	0.394 ± 0.139	$2.820_{(111)}$	**0.0057***	1.439	0.410, 2.468
	C vs 129S2	0.729 ± 0.148	$4.910_{(111)}$	**< 0.0001*****	2.666	1.533, 3.799
	B6N vs 129S2	0.335 ± 0.142	$2.362_{(111)}$	**0.0199***	1.227	0.185, 2.270
Trial 3	C vs B6N	0.060 ± 0.137	$0.442_{(111)}$	0.6595	0.221	-0.772, 1.215
	C vs 129S2	0.333 ± 0.150	$2.222_{(111)}$	0.0283	1.217	0.119, 2.315
	B6N vs 129S2	0.272 ± 0.156	$1.746_{(111)}$	0.0836	0.996	-0.142, 2.133
Trial 4	C vs B6N	-0.031 ± 0.137	$-0.242_{(111)}$	0.8093	-0.121	-1.113, 0.871
	C vs 129S2	0.167 ± 0.155	$1.080_{(111)}$	0.2825	0.613	-0.514, 1.740
	B6N vs 129S2	0.201 ± 0.164	$1.227_{(111)}$	0.2224	0.734	-0.455, 1.923
(b)Risk assessment						
Trial 1	C vs B6N	1.561 ± 0.128	$12.211_{(432)}$	**< 0.0001*****	4.195	3.463, 4.926
	C vs 129S2	0.340 ± 0.123	$2.754_{(432)}$	**0.0061***	0.914	0.259, 1.569
	B6N vs 129S2	-1.221 ± 0.120	$-10.181_{(432)}$	**< 0.0001*****	-3.281	-3.951, -2.160
Trial 2	C vs B6N	0.991 ± 0.095	$10.389_{(432)}$	**< 0.0001*****	2.661	2.127, 3.195
	C vs 129S2	0.028 ± 0.116	$0.244_{(432)}$	0.8073	0.076	-0.535, 0.687
	B6N vs 129S2	-0.962 ± 0.096	$-10.047_{(432)}$	**< 0.0001*****	-2.586	-3.120, 02.051
Trial 3	C vs B6N	0.250 ± 0.078	$3.215_{(432)}$	**0.0014***	0.671	0.259, 1.085
	C vs 129S2	-0.109 ± 0.109	$-0.999_{(432)}$	0.3185	-0.291	-0.866, 0.283
	B6N vs 129S2	-0.359 ± 0.087	$-4.129_{(432)}$	**< 0.0001*****	-0.963	-1.426, -0.500
Trial 4	C vs B6N	0.156 ± 0.067	$2.360_{(432)}$	0.0187	0.419	0.069, 0.769
	C vs 129S2	0.045 ± 0.104	$-0.437_{(432)}$	0.6624	-0.122	-0.670, 0.426
	B6N vs 129S2	-0.201 ± 0.084	$-2.386_{(432)}$	0.0175	-0.541	-0.988, -0.094
(c)Arousal						
Trial 1	C vs B6N	-0.151 ± 0.104	$-1.460_{(111)}$	0.1472	-0.347	-0.819, 0.126
	C vs 129S2	0.255 ± 0.103	$2.472_{(111)}$	**0.0149***	0.584	0.109, 1.059
	B6N vs 129S2	0.406 ± 0.106	$3.847_{(111)}$	**0.0002****	0.931	0.436, 1.426
Trial 2	C vs B6N	0.226 ± 0.111	$2.042_{(111)}$	0.0435	0.517	0.011, 1.024
	C vs 129S2	0.211 ± 0.111	$1.908_{(111)}$	0.0590	0.483	-0.028, 0.989
	B6N vs 129S2	-0.015 ± 0.113	$-0.133_{(111)}$	0.8941	-0.034	-0.546, 0.477
Trial 3	C vs B6N	0.058 ± 0.123	$0.471_{(111)}$	0.6389	0.132	-0.424, 0.689

	C vs 129S2	0.014 ± 0.121	$0.118_{(111)}$	0.9062	0.032	-0.514, 0.580
	B6N vs 129S2	-0.434 ± 0.124	$-0.350_{(111)}$	0.7269	-0.099	-0.663, 0.464
Trial 4	C vs B6N	-0.196 ± 0.135	$-1.458_{(111)}$	0.1477	-0.450	-1.064, 0.165
	C vs 129S2	-0.125 ± 0.134	$-0.933_{(111)}$	0.3528	-0.286	-0.894, 0.322
	B6N vs 129S2	0.072 ± 0.136	$0.525_{(111)}$	0.6005	0.164	-0.455, 0.784
(d)Exploration						
Trial 1	C vs B6N	-0.437 ± 0.071	$-6.114_{(111)}$	**< 0.0001*****	-1.451	-1.960, -0.943
	C vs 129S2	0.017 ± 0.070	$0.247_{(111)}$	0.8050	0.057	-0.401, 0.515
	B6N vs 129S2	0.454 ± 0.071	$6.416_{(111)}$	**< 0.0001*****	1.509	1.001, 2.016
Trial 2	C vs B6N	-0.531 ± 0.090	$-5.915_{(111)}$	**< 0.0001*****	-1.764	-2.401, -1.128
	C vs 129S2	-0.071 ± 0.096	$-0.742_{(111)}$	0.4599	-0.237	-0.870, 0.396
	B6N vs 129S2	0.460 ± 0.101	$4.554_{(111)}$	**< 0.0001*****	1.528	0.833, 2.223
Trial 3	C vs B6N	-0.399 ± 0.107	$-3.733_{(111)}$	**0.0003****	-1.324	-2.049, -0.599
	C vs 129S2	0.175 ± 0.119	$1.468_{(111)}$	0.1450	0.580	-0.207, 1.367
	B6N vs 129S2	0.573 ± 0.124	$4.626_{(111)}$	**< 0.0001*****	1.904	1.050, 2.758
Trial 4	C vs B6N	-0.169 ± 0.093	$-1.807_{(111)}$	0.0736	-0.561	-1.180, 0.058
	B6N vs 129S2	0.428 ± 0.119	$3.587_{(111)}$	**0.0005***	1.423	0.614, 2.231
(e)Locomotion						
Rank transformed						
Trial 1	C vs B6N	-241.65 ± 17.2	$-14.065_{(111)}$	**< 0.0001*****	-2.763	-3.299, -2.228
	C vs 129S2	85.82 ± 20.2	$4.240_{(111)}$	**< 0.0001*****	0.981	0.505, 1.458
	B6N vs 129S2	327.47 ±17.4	$18.809_{(111)}$	**< 0.0001*****	3.745	3.109, 4.380
Trial 2	C vs B6N	-116.75 ± 18.1	$-6.456_{(111)}$	**< 0.0001*****	-1.335	-1.782, -0.889
	C vs 129S2	101.16 ± 21.0	$4.814_{(111)}$	**< 0.0001*****	1.157	0.656, 1.657
	B6N vs 129S2	217.91 ± 18.3	$11.904_{(111)}$	**< 0.0001*****	2.492	1.961, 3.023
Trial 3	C vs B6N	-68.03 ± 19.5	$-3.492_{(111)}$	**0.0007****	-0.778	-1.231, -0.325
	C vs 129S2	131.53 ± 22.1	$5.960_{(111)}$	**< 0.0001*****	1.504	0.965, 2.043
	B6N vs 129S2	199.55 ± 19.6	$10.170_{(111)}$	**< 0.0001*****	2.282	1.744, 2.820
Trial 4	C vs B6N	-6.82 ± 21.1	$-0.324_{(111)}$	0.7467	-0.078	-0.555, 0.399
	C vs 129S2	159.23 ± 23.5	$6.786_{(111)}$	**< 0.0001*****	1.824	1.239, 2.410
	B6N vs 129S2	166.35 ± 21.2	$7.834_{(111)}$	**< 0.0001*****	1.902	1.359, 2.446
(f)pCORT						
Time 1	C vs B6N	0.13 ± 0.06	$2..168_{(79)}$	0.0391***	0.164	0.007, 0.321
	C vs 129S2	-0.02 ± 0.06	$-0.437_{(78)}$	0.6656	-0.031	-0.179, 0.116
	B6N vs 129S2	-0.15 ± 0.06	$-2.436_{(77)}$	0.0217***	-0.196	-0.363, -0.029
Time 2	C vs B6N	-0.10 ± 0.22	$-0.441_{(79)}$	0.6629	-0.123	-0.694, 0.448
	C vs 129S2	-0.87 ± 0.32	$-2.750_{(78)}$	0.0104***	-1.102	-1.937, -0.267
	B6N vs 129S2	-0.78 ± 0.27	$-2.857_{(77)}$	**0.0081*****	-0.979	-1.694, -0.264

Time 3	C vs B6N	0.34 ± 0.10	$3.253_{(79)}$	**0.0030*****	0.424	0.151, 0.698
	C vs 129S2	0.08 ± 0.08	$0.959_{(78)}$	0.3458	0.102	-0.116, 0.319
	B6N vs 129S2	-0.25 ± 0.09	$-2.652_{(77)}$	0.0130***	-0.323	-0.575, -0.070

Table S6. *Post hoc* comparisons between experimenters of the estimated marginal means on each trial, or averaged over trials (in case of a main effect of experimenter), per behavioral dimension. Adjusted $\alpha = 0.025321$ for comparison between experimenters on trials 1 and 4, $P < 0.05$ for trials 2 and 3. Significant contrasts are highlighted in bold.

Dimension		Estimate ± SEM	$t_{(df)}$	*P*	Cohen s *d*	Lower and upper limits of 95% CI
(a)Avoidance						
Overall	Exp. A vs Exp. B	-0.233 ± 0.095	$-2.446_{(111)}$	**0.0160**	-0.380	-0.565, -0.196
(b)Risk assessment						
Trial 1	Exp. A vs Exp. B	0.603 ± 0.101	$5.968_{(432)}$	**< 0.0001**	1.621	1.076, 2.166
Trial 2	Exp. A vs Exp. B	0.047 ± 0.084	$0.562_{(432)}$	0.5747	0.127	-0.317, 0.570
Trial 3	Exp. A vs Exp. B	0.075 ± 0.075	$0.996_{(432)}$	0.3196	0.201	-0.196, 0.598
Trial 4	Exp. A vs Exp. B	0.070 ± 0.070	$0.990_{(432)}$	0.3226	0.187	-0.184, 0.559
(c)Arousal						
Overall	Exp. A vs Exp. B	0.222 ± 0.063	$3.506_{(111)}$	**0.0007**	0.409	0.224, 0.5
(d)Exploration						
Trial 1	Exp. A vs Exp. B	0.499 ± 0.058	$8.560_{(111)}$	**< 0.0001**	1.660	1.220, 2.100
Trial 2	Exp. A vs Exp. B	0.765 ± 0.078	$9.785_{(111)}$	**< 0.0001**	2.540	1.930, 3.160
Trial 3	Exp. A vs Exp. B	0.582 ± 0.095	$6.107_{(111)}$	**< 0.0001**	1.930	1.260, 2.610
Trial 4	Exp. A vs Exp. B	0.627 ± 0.089	$7.021_{(111)}$	**< 0.0001**	2.080	1.430, 2.730
(e)Locomotion						
Exp. A vs Exp. B	C	51.40± 22.3	$2.302_{(111)}$	**0.0232**	0.588	0.076, 1.100
	B6N	-5.26 ± 17.6	$-0.298_{(111)}$	0.7661	-0.060	-0.460, 0.340
	129S2	27.57 ± 22.7	$1.217_{(111)}$	0.2261	0.315	-0.200, 0.831

Table S7. Cluster differences: Effects of different explanatory variables for each behavioral dimension/pCORT.
Behavioral dimensions: analyzed with GLMMs using a 2 (cluster) x 4 (trials) mixed factorial design. Cluster, trial and their interaction were included as fixed predictors, while mouse identity (ID) and mouse slope were included as random factors.
pCORT: LMM using a 2 (cluster) x 3 (sampling moment) mixed factorial design. Cluster, sampling moment and their interaction were included as fixed predictors, mouse ID and slope were included as random factors. Significant effects are highlighted in bold.

Dimension	Explanatory variables	F	Df	P
(a)Avoidance				
	Cluster C	0.12	1, 115	0.7334
	Trial T	10.27	3, 341	**< 0.0001**
	C * T	45.32	3, 341 d	**< 0.0001**
	Mouse identity[r]	211.00 [b]	11	**< 0.0001**
	Mouse slope[r]	7.39 [b]	13	0.0248
(b)Risk assessment				
glmmTMB	Cluster C[a]	0.06[b]	1	0.8060
	Trial T[a]	417.49 [b]	3	**< 0.0001**
	C * T[a]	4.99 [b]	3	0.1722
	Mouse identity[r]	89.64 [b]	11	**< 0.0001**
	Mouse slope[r]	107.83 [b]	13	**< 0.0001**
(c)Arousal				
	Cluster C	0.72	1, 115	0.3960
	Trial T	18.41	3, 341	**< 0.0001**
	C * T	8.59	3, 341	**< 0.0001**
	Mouse identity[r]	36.67 [b]	11	**< 0.0001**
	Mouse slope[r]	7.90 [b]	13	**0.0192**
(d)Exploration				
	Cluster C	0.35	1, 115	0.5561
	Trial T	59.73	3, 341	**< 0.0001**
	C * T	12.42	3, 341	**< 0.0001**
	Mouse identity[r]	289.90 [b]	11	**< 0.0001**
	Mouse slope[r]	7.60 [b]	13	**0.0224**
(e)Locomotion				
rank transformed	Cluster C	1.59	1, 115	0.1923
	Trial T	11.35	3, 341	**0.0010**
	C * T	1.21	3, 341	0.3046
	Mouse identity[r]	184.36 [b]	11	**< 0.0001**
	Mouse slope[r]	Na	na	na
(f)pCORT	Cluster C	2.73	1, 115	0.2429
	Time T	224.8	2, 213	**< 0.0001**
	C*T	2.11	2, 213	0.0880
	Mouse identity[r]	4.51 [b]	8	**0.0336**
	Mouse slope[r]	0.00 [b]	10	1.0000

[r]Mouse identity and mouse slope were included as random factors; the statistical significance of these factors was calculated by likelihood-ratio tests and thus [b]Chi-square values are reported.
[a]Analyses conducted with glmmTMB: [b]main and interaction effects reported with Chi Square values.
Na = model did not converge.

Table S8. *Post hoc* within cluster comparisons of the estimated marginal means between trials 1 and 4 (behavioral dimensions) or sampling moments (pCORT). Behavioral dimensions: adjusted α = 0.025321 for comparison between trial 1 and 4. pCORT: adjusted α = 0.025321 for comparison between sampling moments. Significant contrasts are highlighted in bold.

Dimension		Estimate ± SEM	$t_{(df)}$	*P*	Cohens *d*	Lower and upper limits of 95% CI
(a)Avoidance						
Trial 1 vs 4	A	0.948 ± 0.089	$10.636_{(341)}$	**< 0.0001**	1.558	1.247, 1.869
Trial 1 vs 4	B	-0.608 ± 0.116	$-5.252_{(341)}$	**< 0.0001**	-1.000	-1.382, -0.618
(c)Arousal						
Trial 1 vs 4	A	-0.260 ± 0.077	$-3.382_{(115)}$	**0.0010**	-0.641	-1.026, -0.256
Trial 1 vs 4	B	-0.762 ± 0.100	$-7.621_{(115)}$	**< 0.0001**	-1.875	-2.420, -1.329
(d)Exploration						
Trial 1 vs 4	A	-0.853 ± 0.064	$-13.420_{(341)}$	**< 0.0001**	-2.494	-2.905, -2.083
Trial 1 vs 4	B	-0.323 ± 0.070	$-4.659_{(341)}$	**< 0.0001**	-0.935	-1.348, -0.539
(f)pCORT						
Time 1 vs 2	Overall	-1.579 ± 0.098	$-16.038_{(115)}$	**< 0.0001**	-8.707	-10.720, -7.141
Time 1 vs 3	Overall	-0.145 ± 0.037	$-3.959_{(115)}$	**0.0001**	-0.801	-1.220, -0.387
Time 2 vs 3	Overall	1.433 ± 0.100	$14.430_{(115)}$	**< 0.0001**	7.905	6.400, 9.409

Table S9. *Post hoc* between cluster comparisons of the estimated marginal means on each trial for avoidance behavior, arousal and exploration (adjusted α = 0.025321 for trials 1 and 4, α = 0.05 for trials 2 and 3). Significant contrasts are highlighted in bold.

Dimension		Estimate ± SEM	$t_{(df)}$	P	Cohen's *d*	
(a)Avoidance						
A vs B	Trial 1	1.131 ± 0.164	$7.885_{(115)}$	**< 0.0001**	1.860	1.272, 2.448
	Trial 2	0.420 ± 0.132	$3.175_{(115)}$	**0.0019**	0.690	0.250, 1.130
	Trial 3	-0.152 ± 0.134	$-1.133_{(115)}$	0.2594	-0.250	-0.688, 0.188
	Trial 4	-0.425 ± 0.129	$-3.284_{(115)}$	**0.0014**	-0.698	-1.129, -0.267
(c)Arousal						
A vs B	Trial 1	0.162 ± 0.090	$1.792_{(115)}$	0.0757	0.400	-0.045, 0.845
	Trial 2	0.047 ± 0.094	$0.493_{(115)}$	0.6233	0.115	-0.346, 0.576
	Trial 3	-0.345 ± 0.100	$-3.441_{(115)}$	**0.0008**	-0.849	-1.350, -0.348
	Trial 4	-0.339 ± 0.107	$-3.163_{(115)}$	**0.0020**	-0.834	-1.367, -0.300
(d)Exploration						
A vs B	Trial 1	-0.166 ± 0.096	$-1.729_{(115)}$	0.0864	-0.485	-1.044, 0.074
	Trial 2	-0.068 ± 0.102	$-0.663_{(115)}$	0.5083	-0.199	-0.792, 0.395
	Trial 3	0.221 ± 0.110	$2.020_{(115)}$	**0.0458**	0.647	0.007, 1.287
	Trial 4	0.364 ± 0.117	$3.115_{(115)}$	**0.0023**	1.065	0.374, 1.757

Table S10. Multiple comparisons: Overview of Dunn-Sidak corrected values for α in *post hoc* tests.

Results section		GLMM effect	Post hoc comparisons/contrasts	γ	Adjusted α
3.1. Strain analyses	Behavior	Strain	C vs B6; C vs 129S2; B6 vs 129S2	2	0.025321
	Behavior	Trial	Trial 1 vs Trial 4	1	0.05
	Behavior	Strain x Trial (T)	**C-T1** vs C-T4; C-T1 vs B6-T1; C-T1 vs 129S2-T1;	3	0.01692
			B6-T1 vs B6-T4; C-T1 vs B6-T1; B6-T1 vs 129S2-T1	3	0.01692

			129S2-T1 vs 129S2-T4; C-T1 vs 129S2-T1; B6-T1 vs 129S2-T1	3	0.01692
			C-T2 vs B6-T2; C-T2 vs 129S2-T2; B6-T2 vs 129S2-T2	2	0.025321
			C-T3 vs B6-T3; C-T3 vs 129S2-T3; B6-T3 vs 129S2-T3	2	0.025321
			C-T1 vs **C-T4**; C-T4 vs B6-T4; C-T4 vs 129S2-T4;		
			B6-T1 vs **B6-T4**; C-T4 vs B6-T4; B6-T4 vs 129S2-T4		
			129S2-T1 vs **129S2-T4**; C-T4 vs 129S2-T4; B6-T4 vs 129S2-T4		
	Behavior	Experimenter (E) x Trial (T)	E1-T1 vs E1-T4; E1-T1 vs E2-T1; E1-T4 vs E2-T4	2	0.025321
			E2-T1 vs E2-T4; E1-T1 vs E2-T1; E1-T4 vs E2-T4	2	0.025321
			E1-T2 vs E2-T2	1	0.05
			E1-T3 vs E2-T3	1	0.05
		Experimenter (E) x Strain	E1-C vs E2-C; E1-B6 vs E2-B6; E1-129S2 vs E2 – 129S2	1	0.05
	Corticosterone	Strain x Sampling moment (S)	**C-S1** vs C-S2; **C-S1** vs C-S3; C-S2 vs C-S3; B6-S1 vs B6-S2; B6-S1 vs B6-S3; B6-S2 vs B6-S3; ; 129S2-S1 vs 129S2-S2; 129S2-S1 vs 129S2-S3; 129S2-S2 vs 129S2-S3; **C-S1** vs B6-S1; **C-S1** vs 129S2-S1; B6-S1 vs 129-S1; C-S2 vs B6-S2; C-S2 vs 129S2-S2; B6-S2 vs 129-S2; C-S3 vs B6-S3; C-S1 vs 129S2-S3; B6-S3 vs 129-S3;	4	0.012741
3.3. Cluster analyses	Behavior	Cluster (A/B) x Trial (T)	A-T1 vs A-T4; B-T1 vs B-T4; A-T1 vs B-T1; A-T4 vs B-T4	2	0.025321
			A-T2 vs B-T2; A-T3 vs B-T3	1	0.05
	Corticosterone	Sampling moment (S)	S1 vs S2; S1 vs S3; S2 vs S3	2	0.025321
3.6 Clusters within strains	Behavior	Cluster (A/B) X Trial (T)	A-T1 vs A-T4; B-T1 vs B-T4; A-T1 vs B-T1; A-T4 vs B-T4	2	0.025321

			A-T2 vs B-T2; A-T3 vs B-T3	1	0.05
		Trial (T)	T1 vs T4	1	0.05
	Corticosterone	Sampling moment (S)	S1 vs S2; S1 vs S3; S2 vs S3	2	0.025321

Table S11. Cluster (A/B) differences within strains C, B6N and 129S2: Effects of different explanatory variables for each behavioral dimension/pCORT. Behavioral dimensions were analyzed with generalized linear mixed models (GLMMs) using a 2 (cluster) x 4 (trials) mixed factorial design. Cluster, trial and their interactions were included as fixed predictors. Mouse identity (ID)was included as random factor.
pCORT: linear mixed model (LMM) using a 2 (cluster) x 3 (sampling moment) mixed factorial design. Cluster, sampling moment were included as fixed predictors, including their interaction. Mouse identity was included as random factor. Significant effects ($P < 0.05$) are highlighted in bold.

Strain: C					
	Explanatory variables	F	df	P	
Avoidance	Cluster C	0.71	1, 38	0.4043	
	Trial T	26.65	3, 114	**< 0.0001**	
	C * T	9.25	3, 114	**< 0.0001**	
Arousal	Cluster C	4.59	1, 38	**0.0386**	
	Trial T	7.13	3, 114	**0.0002**	
	C * T	5.10	3, 114	**0.0024**	
Risk assessment	Cluster C	0.20	1	0.6590	
glmmTMB	Trial T	413.61	3	**< 0.0001**	
	C * T	5.08	3	0.1655	
Exploration	Cluster C	0.02	1, 38	0.8875	
	Trial T	61.74	3, 114	**< 0.0001**	
	C * T	1.18	3, 114	0.3195	
Locomotion	Cluster C	0.75	1, 38	0.3906	
	Trial T	7.31	3, 114	**0.0002**	
	C * T	0.66	3, 114	0.5747	
Corticosterone	Cluster C	3.47	1, 38	0.0701	
	Trial T	49.01	2, 73	**< 0.0001**	
	C * T	1.51	2, 73	0.2259	
Strain: B6N					
Avoidance	Cluster C	0.38	1, 37	0.7519	
	Trial T	4.09	3, 109	**0.0086**	
	C * T	13.41	3, 109	**< 0.0001**	
Arousal	Cluster C	3.72	1, 37	0.0615	
	Trial T	4.95	3, 109	**0.0029**	
	C * T	4.38	3, 109	**0.0059**	
Risk assessment	Cluster C	0.65	1	0.4200	
glmmTMB	Trial T	178.24	3	**< 0.0001**	
	C * T	0.55	3	0.9070	
Exploration	Cluster C	0.02	1, 37	0.8909	
	Trial T	38.04	3, 109	**< 0.0001**	
	C * T	7.11	3, 109	**0.0002**	
Locomotion	Cluster C	6.30	1, 37	**0.0165**	

	Trial T	63.36	3, 109	**< 0.0001**	
	C * T	0.27	3, 109	0.8473	
Corticosterone	Cluster C	0.18	1, 37	0.6689	
	Trial T	154.34	2,67	**< 0.0001**	
	C * T	2.89	2, 67	0.0627	
Strain: 129S2					
Avoidance	Cluster C	1.06	1, 36	0.3101	
	Trial T	0.88	3, 106	0.4533	
	C * T	21.53	3, 106	**< 0.0001**	
Arousal	Cluster C	1.32	1, 36	0.2585	
	Trial T	17.89	3, 106	**<0.0001**	
	C * T	0.08	3, 106	0.9719	
Risk assessment	Cluster C	1.04	1, 36	0.3071	
	Trial T	553.84	3, 106	**< 0.0001**	
	C * T	4.69	3, 106	0.1958	
Exploration	Cluster C	0.67	1, 36	0.4162	
	Trial T	13.10	3, 106	**< 0.0001**	
	C * T	7.51	3, 106	**0.0001**	
Locomotion	Cluster C	1.43	1, 36	0.2396	
	Trial T	0.68	3, 106	0.5673	
	C * T	0.32	3, 106	0.8074	
Corticosterone	Cluster C	0.34	1, 36	0.2599	
	Time T	143.29	2, 65	**< 0.0001**	
	C * T	0.41	2, 65	0.8197	

Table S12. *Post hoc* comparisons of the estimated marginal means between and/or within clusters, for each strain (C/B6N/129S2) separately.
Within cluster *post hoc* comparisons between trials 1 and 4 (behavioral dimensions) or sampling moments (pCORT): Behavioral dimensions: adjusted $\alpha = 0.025321$ for comparison between trial 1 and 4. pCORT: adjusted $\alpha = 0.025321$ for comparison between sampling moments.
Post hoc between cluster comparisons on each trial: Behavioral dimensions: (adjusted $\alpha = 0.025321$ for trials 1 and 4, $\alpha = 0.05$ for trials 2 and 3). Significant contrasts are highlighted in bold.

Strain: C		Estimate ± SEM	$t_{(df)}$	P
Avoidance				
Cluster A	Trial 1 vs 4	1.241 ± 0.128	$9.661_{(114)}$	**< 0.0001**
Cluster B	Trial 1 vs 4	-0.236 ± 0.279	$-0.847_{(114)}$	0.3985
Trial 1				
	A vs B	0.944 ± 0.323	$2.919_{(38)}$	**0.0059**
Trial 2				
	A vs B	0.169 ± 0.245	$0.692_{(38)}$	0.4934
Trial 3				
	A vs B	-0.275 ± 0.221	$-1.245_{(38)}$	0.2208
Trial 4				
	A vs B	-0.534 ± 0.220	$-2.424_{(38)}$	**0.0202**
Risk assessment				

Trial main				
	Trial 1 vs 4	2.316 ± 0.151	$15.267_{(146)}$	**< 0.0001**
Arousal				
Cluster A				
	Trial 1 vs 4	-0.234 ± 0.099	$-2.376_{(114)}$	**0.0192**
Cluster B				
	Trial 1 vs 4	-0.854 ± 0.214	$-3.992_{(114)}$	**0.0001**
Trial 1				
	A vs B	0.159 ± 0.185	$0.860_{(38)}$	0.3951
Trial 2				
	A vs B	-0.036 ± 0.185	$-0.196_{(38)}$	0.8455
Trial 3				
	A vs B	-0.659 ± 0.185	$-3.556_{(38)}$	**0.0010**
Trial 4				
	A vs B	-0.460 ± 0.185	$-2.486_{(38)}$	**0.0174**
Exploration				
Trial main				
	Trial 1 vs 4	-0.685 ± 0.080	$-8.597_{(114)}$	**< 0.0001**
Locomotion				
Trial main				
	Trial 1 vs 4	-0.414 ± 0.108	$-3.807_{(114)}$	**0.0002**
Corticosterone				
Time main				
	Time 1 vs 2	-1.298 ± 0.203	$-6.394_{(73)}$	**< 0.0001**
	Time 1 vs 3	-0.507 ± 0.177	$-2.863_{(73)}$	**0.0055**
	Time 2 vs 3	0.791 ± 0.160	$4.938_{(73)}$	**< 0.0001**
Strain: B6N		Estimate ± SEM	$t_{(df)}$	P
Avoidance				
Cluster A				
	Trial 1 vs 4	0.644 ± 0.116	$5.573_{(109)}$	< 0.0001
Cluster B				
	Trial 1 vs 4	-0.307 ± 0.127	$-2.383_{(109)}$	**0.0189**
Trial 1				
	A vs B	0.714 ± 0.232	$3.079_{(37)}$	**0.0039**
Trial 2				
	A vs B	0.234 ± 0.231	$1.015_{(37)}$	0.3166
Trial 3				
	A vs B	-0.209 ± 0.231	$-0.905_{(37)}$	0.3715
Trial 4				
	A vs B	-0.233 ± 0.231	$-1.011_{(37)}$	0.3187
Risk assessment				
Trial main				
	Trial 1 vs 4	0.805 ± 0.072	$11.183_{(151)}$	**< 0.0001**
Arousal				
Cluster A				
	Trial 1 vs 4	-0.118 ± 0.134	$-0.885_{(109)}$	0.3779
Cluster B				
	Trial 1 vs 4	-0.771 ± 0.190	$-4.054_{(109)}$	0.0001
Trial 1				

	A vs B	0.180 ± 0.174	$1.038_{(37)}$	0.3060
Trial 2				
	A vs B	0.040 ± 0.171	$0.237_{(37)}$	0.8143
Trial 3				
	A vs B	-0.172 ± 0.172	$-2.805_{(37)}$	**0.0080**
Trial 4				
	A vs B	-0.427 ± 0.171	$-2.770_{(37)}$	**0.0087**
Exploration				
Cluster A				
	Trial 1 vs 4	0.822 ± 0.079	$10.331_{(109)}$	**< 0.0001**
Cluster B				
	Trial 1 vs 4	0.781 ± 0.088	$8.880_{(109)}$	**< 0.0001**
Trial 1				
	A vs B	-0.275 ± 0.171	$-1.610_{(37)}$	0.1158
Trial 2				
	A vs B	-0.067 ± 0.170	$-0.395_{(37)}$	0.6949
Trial 3				
	A vs B	0.197 ± 0.170	$1.156_{(37)}$	0.2551
Trial 4				
	A vs B	0.256 ± 0.170	$1.605_{(37)}$	0.1406
Locomotion				
Cluster main				
	A vs B	0.177 ± 0.076	$2.323_{(37)}$	**0.0258**
Trial main				
	Trial 1 vs 4	0.801 ± 0.059	$13.517_{(109)}$	**< 0.0001**
Corticosterone				
Trial main				
	Time 1 vs 2	-2.079 ± 0.158	$-13.185_{(67)}$	**< 0.0001**
	Time 1 vs 3	-0.197 ± 0.196	$-1.008_{(67)}$	0.3169
	Time 2 vs 3	1.882 ± 13.505	$13.505_{(67)}$	**< 0.0001**
129S2		Estimate ± SEM	$t_{(df)}$	P
Avoidance				
Cluster A				
	Trial 1 vs 4	0.579 ± 0.136	$4.258_{(106)}$	**< 0.0001**
	Trial 1 vs 4	-0.884 ± 0.148	$-5.979_{(106)}$	**< 0.0001**
Trial 1				
	A vs B	0.394 ± 0.266	$1.483_{(36)}$	0.1467
Trial 2				
	A vs B	0.175 ± 0.266	$0.658_{(36)}$	0.5147
Trial 3				
	A vs B	-0.440 ± 0.265	$-1.660_{(36)}$	0.1057
Trial 4				
	A vs B	-1.070 ± 0.265	$-4.038_{(36)}$	**0.0003**
Risk assessment				
Trial main				
	Trial 1 vs 4	1.821 ± 0.093	$19.470_{(144)}$	**< 0.0001**
Arousal				
Trial main				

	Trial 1 vs 4	-0.708 ± 0.120	$-5.920_{(106)}$	**< 0.0001**
Exploration				
Cluster A				
	Trial 1 vs 4	-1.172 ± 0.178	$-6.575_{(106)}$	**< 0.0001**
Cluster B				
	Trial 1 vs 4	-0.333 ± 0.100	$-3.319_{(106)}$	**0.0012**
Trial 1				
	A vs B	-0.250 ± 0.230	$-1.090_{(36)}$	0.2827
Trial 2				
	A vs B	-0.095 ± 0.230	$-0.412_{(36)}$	0.6829
Trial 3				
	A vs B	0.395 ± 0.229	$1.722_{(36)}$	0.0937
Trial 4				
	A vs B	0.588 ± 0.229	$2.566_{(36)}$	**0.0146**
Corticosterone				
Time main				
	Time 1 vs 2	-1.813 ± 0.150	$-12.074_{(65)}$	**< 0.0001**
	Time 1 vs 3	-0.273 ± 0.190	$-1.437_{(65)}$	0.1556
	Time 2 vs 3	1.540 ± 0.162	$9.532_{(65)}$	**< 0.0001**