Journal of Neuroscience Methods 343 (2020) 108810

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

An individual based, multidimensional approach to identify emotional reactivity profiles in inbred mice

Marloes H. van der Goot[a,*], Hetty Boleij[a], Jan van den Broek[b], Amber R. Salomons[a], Saskia S. Arndt[a], Hein A. van Lith[a,c]

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

ARTICLE INFO

Keywords:
Phenotypic variation
Inbred mice
Behavioral profile
Habituation
Sensitization
Cluster analysis

ABSTRACT

Background: Despite extensive environmental standardization and the use of genetically and microbiologically defined mice of similar age and sex, individuals of the same mouse inbred strain commonly differ in quantitative traits. This is a major issue as it affects the quality of experimental results. Standard analysis practices summarize numerical data by means and associated measures of dispersion, while individual values are ignored. Perhaps taking individual values into account in statistical analysis may improve the quality of results.
New method: The present study re-inspected existing data on emotional reactivity profiles in 125 BALB/cJ and 129 mice, which displayed contrasting patterns of habituation and sensitization when repeatedly exposed to a novel environment (modified Hole Board). Behaviors were re-analyzed on an individual level, using a multivariate approach, in order to explore whether this yielded new information regarding subtypes of response, and their expression between and within strains.
Results: Clustering individual mice across multiple behavioral dimensions identified two response profiles: a habituation and a sensitization cluster.
Comparison with existing method(s): These retrospect analyses identified habituation and sensitization profiles that were similar to those observed in the original data but also yielded new information such as a more pronounced sensitization response. Also, it allowed for the identification of individuals that deviated from the predominant response profile within a strain.
Conclusions: The present approach allows for the behavioral characterization of experimental animals on an individual level and as such provides a valuable contribution to existing approaches that take individual variation into account in statistical analysis.

1. Introduction

Most animal studies for research and other scientific purposes use laboratory mice; In the EU for example they account for more than half of the vertebrate experimental animals (Dutta and Sengupta, 2016). Furthermore, approximately 80 % of the (published) laboratory mouse studies worldwide are conducted with inbred strains (Festing, 2014). A major issue however is that, despite the use of genetically and microbiologically defined laboratory mice and extensive environmental standardization, considerable differences in quantitative biological traits – like behavior – between individual animals of the same inbred strain, age and sex are still found (Loos et al., 2015; Jensen et al., 2016; Einat et al., 2018). In fact, inbred strains, when compared to outbred stocks, display similar trait variability (Tuttle et al., 2018). Apparently, there is another component that contributes to the individual (behavioral) phenotype in inbred mice, one that is not controlled for with environmental and genotypic standardization (Beynen, 1991; Beynen et al., 2001; Gärtner, 2012).

The exact constitution of this so-called ‘third component’ (Gärtner, 2012) remains unclear, although many sources of variation have been identified, ranging from nuclear genetic, epigenetic, mitochondrial genetic and environmental factors (e.g. Crabbe et al., 1999; Freund

Abbreviations: mHB, modified hole board; CVI, clustering validity index; GLMM, generalized linear mixed models

[*] Corresponding author at: Department Population Health Sciences, Unit Animals in Science and Society, Faculty of Veterinary Medicine, Utrecht University, Building Nieuw Gildestein, Yalelaan 2, 3584 CM Utrecht, the Netherlands.

E-mail address: m.h.vandergoot@uu.nl (M.H. van der Goot).

https://doi.org/10.1016/j.jneumeth.2020.108810
Received 16 March 2020; Received in revised form 12 June 2020; Accepted 12 June 2020
Available online 20 June 2020

et al., 2013; Loos et al., 2015) to variation in the gut-microbiome (Burokas et al., 2017; Sandhu et al., 2017). These findings indicate that the existence of phenotypic difference between individuals of the same inbred strain which are kept under standardized husbandry practices and are subject to standardized experimental protocols, is the result of complex interactions between the aforementioned sources of variation. As a consequence, even individuals that share a genetic background differ in their behavior or response to some extent (Koolhaas et al., 2010).

A basic rule of good design of animal experiments is that all variables should be controlled except that due to the treatment (Festing, 2011; Festing et al., 2016). From a laboratory animal science perspective, this complex interaction between different sources of variation makes it challenging to completely control or eliminate all sources of inter-individual variation in animal experiments. An alternative approach might in turn be to improve control by taking individual phenotypic variation into account in experimental design and statistical analysis, rather than dismissing it as noise (Bello and Renter, 2018; Karp, 2018).

Standard analysis practices however summarize numerical data by means and standard deviation, standard error of the mean, 95 % confidence interval and/or medians with the interquartile range. By presenting the data this way one focuses mainly on the means (or medians) and the associated *P* values from the statistical analyses. Since statistical significance represented by *P* values may not necessarily predicate practical importance, some scientists also emphasize the importance of reporting effects sizes (e.g. Labots et al., 2018; Wahlsten, 2011). In any event, when describing data with means (or medians), measures of dispersion, *P* values and effect sizes, individual values are ignored. If one is able to behaviorally define experimental animals on an individual level and incorporate these findings into the study design and statistical analyses, then this may contribute to the quality of any animal experiment (i.e. not only to the quality of behavioral animal experiments) (Garner, 2005). Further, it may lead to a more accurate estimation of the optimal number of experimental units (often the experimental unit is a single animal) needed for such an experiment.

Incorporating individual variability may be of special importance in preclinical animal models on behavioral disorders and psychopathologies (Armario and Nadal, 2013; Ebner and Singewald, 2017; Einat et al., 2018). In human patients, the susceptibility to develop neuropsychological disorders, and the response to treatment is known to vary greatly between individuals (Einat et al., 2018). As such, Einat et al. (2018) for instance argued that animal models may become more representative and homologous when individual differences are taken into account. Increased knowledge on individual variability of behavior and/or response to treatment in model animals may improve understanding of differential vulnerability to development of disorders or patterns in response to treatment, as well as the neurobiological substrates that characterize these differential responses (Armario and Nadal, 2013; Einat et al., 2018).

What type(s) of characteristics are addressed when defining individual animals however, naturally depends on the research objective. To acquire more meaningful behavioral data several reports on the application of multivariate techniques in the study of exploration and anxiety-related behavior in rodents have been produced. Some of these studies have utilized approaches based on the analysis of transition matrices (e.g. Spruijt and Gispen, 1984; Casarrubea et al., 2009; Spruijt et al., 2014) and T-pattern analysis (Magnusson, 2000; Casarrubea et al., 2014, 2015). The work of these authors emphasizes that functionality of individual behaviors can only be fully understood when placed in the (temporal/sequential) context of other behaviors that are displayed. However, the objective of these approaches is not so much directly related to the assessment of behavior of individual animals, but rather to interpreting and analyzing individual behavioral acts in the context of other behaviors expressed by either the same animal or on average by a group of animals. As such these particular multivariate approaches lie beyond the scope of this study.

In other fields however, particularly behavioral ecology, an increasing number of frameworks have been developed that consider and/or facilitate the analysis of individual variation (e.g. Dingemanse and Dochtermann, 2013; Araya-Ajoy et al., 2015; Allegue et al., 2017; Bushby et al., 2018; Reed et al., 2019; Voelkl and Würbel, 2019). The majority of these approaches rely on multilevel models (i.e. generalized linear mixed models). In these models, different variance components related to individual variation are summarized to single point data. For example, they enable researchers to estimate which amount of variation in the data is related to differences between individual animals (measured as the deviation of individual intercepts from the population intercept, related to animal personality, Réale and Dingemanse 2012).

In some cases however, one may be interested in defining individuals on yet another characteristic: the shape or progression of behavioral (or physiological) response curves (Galatzer-Levy et al., 2013; Reed et al., 2019). When zooming in on individual response curves, one might for example want to assess the extent to which groups of individuals follow the same response over time in a population, and delineate the characteristics of these individuals (Nagin, 1999; Genolini et al., 2015; Galatzer-Levy et al., 2013). In those instances, the evolution of a response (e.g. the increase/decrease of a response) is of interest, rather than the deviation from a population intercept.

This may be of interest in research on behavioral habituation and sensitization in the context of preclinical anxiety research. These two contrasting forms of non-associative learning are viewed 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 (Eisenstein and Eisenstein, 2006). In preclinical anxiety research, successful habituation of anxiety related responses is considered an adaptive emotional response that allows individuals to adapt to environmental challenges (Salomons et al., 2010b; Ohl et al., 2008). In a series of mouse studies, Salomons, Boleij and colleagues assessed whether the opposite of such a response (i.e. a sensitization of anxiety responses) may then reflect a non-adaptive anxiety response, and - ultimately – whether this phenomenon may be employed as a symptom of pathological anxiety in mouse models (Boleij et al., 2012; Salomons et al., 2010a, b; Salomons et al., 2010c, 2013).

In these studies, (sub-)strains of BALB/c and 129 mice were behaviorally characterized by repeated exposure to the modified Hole Board (mHB) test (Labots et al., 2015). The BALB/c strain is the most commonly used mouse inbred strain in animal experimentation (≈46 %; Festing, 2014), whereas the 129 mouse was the most widely used strain in gene targeting experiments (Cook et al., 2002). These strains show distinct contrasting behaviors in tests of anxiety, and are therefore often used in preclinical anxiety studies. In the aforementioned studies, mice from the BALB/cJ strain were characterized by initial high levels of anxiety-related behavior that decreased as trials progressed, while exploratory and locomotor behavior increased over time. This indicated successful habituation to the behavioral test. In contrast, the profile of mice from the 129P3/J strain was characterized by a lack of habituation as initial low levels of anxiety-like behavior increased as trials progressed, while exploration and locomotor activity largely remained stable over time. This indicated a sensitization response to the same experimental set up.

These profiles were based on the (sub-)strain means and medians. Retrospect analyses on these studies however, showed that variation in anxiety-like responses within strains was quite substantial: it was not unusual to find coefficients of variation over 100 % (exemplary variable: percentage of time spent on board; Salomons et al., 2010a). Perhaps the ‘third component’ played a role here as well.

In the present paper we therefore re-inspected the data of these experiments by zooming in on response curves of individual mice, instead of average strain responses. These response curves will be referred to as trajectories from here on, as is common in longitudinal studies

(Genolini et al., 2015). Our objective was to explore the data for subgroups of individual mice, regardless of strain, that displayed similar trajectories across trials, and that consistently grouped together across multiple behavioral dimensions: distinct types of behavioral response profiles. To do this we used a k-means clustering procedure that was specifically designed for grouping of multiple longitudinal response trajectories, kml3d (Genolini et al., 2015). We asked whether this approach would yield new information regarding subtypes of behavioral profiles, and how different profiles were divided across and within strains.

In order to do this, we first summarized the behavioral variables. Anxiety-related behavior is expressed by a combination of behavioral dimensions, such as avoidance (Belzung and Griebel, 2001), risk assessment (Rodgers and Dalvi, 1997), arousal (O'Leary et al., 2013), but also locomotor activity and exploration; the latter acts as counterpart of expressed anxiety (Ohl, 2003; Laarakker et al., 2008; Labots et al., 2016). Moreover, previous research showed that behavioral variables observed in the modified Hole Board can be summarized in five behavioral dimensions: avoidance, risk assessment, arousal, locomotion and exploration (Laarakker et al., 2008, 2011; Labots et al., 2018). It was therefore considered desirable to use so-called composite variables that represent these underlying dimensions rather than single behavioral variables to classify habituation and sensitization patterns.

Hence, in order to assess whether the 'third component' may be present in the habituation and sensitization responses of inbred mice, the yielded composite variables were analyzed across experiments and strains using the k-means clustering procedure by Genolini et al. (2015). The number of different behavioral response profiles that were displayed and how these profiles were expressed within and between inbred strains of mice are described below.

2. Materials and methods

The data in the present paper combined data from five previously published studies (Boleij et al., 2012; Salomons et al., 2010a, b; Salomons et al., 2010c, 2013). The underlying animal experiments all followed the same procedure with respect to animal handling, housing, experimental protocol and ethical permission. These procedures are described below. The experiments also differed in factors such as (sub-) strain, sex, age at behavioral testing, experimenter, animal supplier or housing location. Appendix Table A1 gives an overview of these factors for each study.

2.1. Animals and housing

The experiments were performed on 125 naïve male and female mice of two different mouse inbred strains: BALB/cJ (N = 40; female N = 10) and 129P3/J (N = 53, female N = 10), and four other substrains of the 129-family: 129S2/SvPasCrl (N = 8), 129S2/SvHsd (N = 8), 129 × 1/J (N = 8) and 129P2/OlaHsd (N = 8), all males. For detailed information on stock numbers, supplier, age of testing and sex, see appendix Table A1.

Experiments were conducted at three different locations (see appendix Table A1). In all locations similar housing conditions applied. Animals were housed individually in Macrolon Type II (size 268 × 215 × 141 mm, floor area 370 cm^2) or Macrolon Type II L cages (size: 365 × 207 × 140 mm, floor area 530 cm^2, Techniplast, Milan, Italy) with standard bedding material (autoclaved Aspen Chips, Abedd-Dominik Mayr KEG, Köflach, Austria) and a tissue (KLEENEX® Facial Tissue, Kimberley-Clark Professional BV, Ede, the Netherlands) and cardboard shelter as enrichment. Food (CRM, Expanded, Special Diets Services Witham, England) and water were available *ad libitum.*

All animals were kept in a laboratory animal housing room for a habituation period of 17 days under a reversed 12 h/12 h light/dark cycle (lights off at 6:00) and a radio played constantly as background noise. The mice were handled three times a week during this period by the person who conducted the experiment. Relative humidity was kept at a constant level of 50 % (± 5) with an average room temperature of 22^{0}C (± 2) and a ventilation rate of 15–20 changes/hour.

2.2. Modified hole board

All 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 (Ohl et al., 2001). It is aimed at 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. At the same time, it overcomes the disadvantages of a test battery, by reducing the number of animals, and the time, used for testing. Further it circumvents the possible effect of test order as well as the risk of that the experience of one test carries over to another one (Ohl et al., 2001; Labots et al., 2015). A drawback is that the behavior in the mHB is manually scored for a certain period of time and manual scoring is more laborious compared to an automated scoring system. In addition, automatic scoring allows more data collection. Also, handling and possible influence of the experimenter weighs heavier on the manually scored behavioral outcome compared to an automated procedure.

The mHB paradigm has been described extensively elsewhere (see Labots et al., 2015) and will only be briefly explained here. The apparatus consists of a grey 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 (Fig. 1). The area around the board is divided into 10 rectangles (20 × 15 cm) and 2 squares (20 × 20 cm). In our experiments, this 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 in order to increase the aversive nature of the central (unprotected) area. Light intensity was either 50 lx or 120 lx, depending on the study (see appendix Table A1).

2.3. Experimental protocol

Testing took place in the same room as where the animals were housed, and test equipment was placed in the room prior to arrival of the animals. Testing occurred between 09:00 and 13:00, during the active phase of the animals. Experiments were conducted by four different experimenters, see appendix Table A1. Test procedure was the same across experiments. All mice were tested individually for a total of 20 trials. Each trial lasted 5 min, and mice were tested in a randomized order for 5 consecutive days (4 trials/day). Prior to start of the trial, the home cage was placed next to the mHB. Mice were picked up at the tail

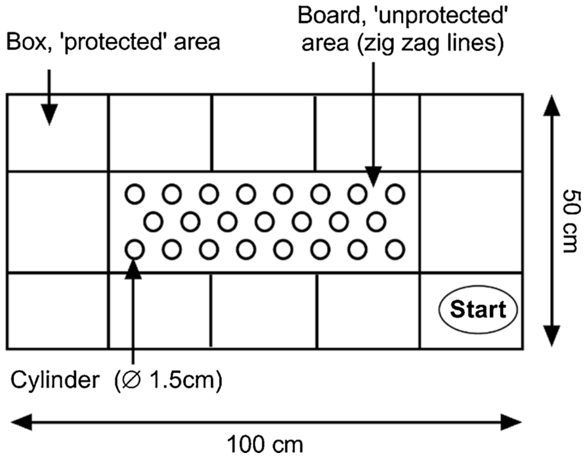

Fig. 1. Schematic overview of the modified Hole Board.

base, transferred from the home cage to the mHB and always placed in the same corner, facing the central board. During the test, mice were allowed to freely explore the mHB-set up. After each trial the mHB was carefully cleaned with water and a damp towel. Behavior was scored live by using the software Observer (Noldus Technology, Wageningen, the Netherlands; for Observer versions per experiment see appendix Table A1). Trials were simultaneously recorded on camera for raw data storage.

2.4. Behavioral dimensions

Behavioral profiles were assessed by scoring behavioral variables listed in appendix Table A2. These behaviors were scored as separate variables during testing. However, as described in the introduction, previous studies have shown that behaviors scored in the mHB can be summarized in five behavioral dimensions: avoidance behavior, risk assessment, arousal, exploration and locomotion (Laarakker et al., 2008, 2011; Labots et al., 2018). In the present manuscript scores on the original variables were therefore combined to these five underlying behavioral dimensions using the procedure described below. All dimensions and corresponding behavioral variables are specified in appendix Table A2.

2.5. Integrated behavioral z-score calculation

Guilloux et al. (2011) proposed the method of integrated behavioral z-scoring as a method for behavioral phenotyping in mice. In this approach behavioral variables that measure different aspects of behavior are normalized and combined to a single score representing that underlying behavioral dimension or motivational system (Labots et al., 2018). Normalization is done by z-score transformation, which assesses the amount of standard deviations each observation is above or below the mean of a reference or control group (Guilloux et al., 2011). The advantage of integrated behavioral z-scores is that they are not constrained by criteria that are demanded by other multivariate approaches like principal component/factor analysis (such a behavioral variable to sample size ratio of at least1:3, Budaev, 2010).

A potential drawback of this approach is that the determination of the reference or control group is not always straightforward, depending on the study design. Control groups may not always be available, for example in studies that directly compare behavior between two inbred strains. This was the case in the experimental studies that were combined for analysis in the present paper (Boleij et al. (2012); Salomons et al., 2010a, b; Salomons et al., 2010c, 2013).

Also, a problem may occur when the control group used for the calculation of the z-scores has a standard deviation of zero. Labots et al. (2018) therefore suggested an improved calculation procedure, in which the combined data of all experimental groups in a study is used as a reference group. A standard deviation of zero in a pooled dataset would imply that there is no variance in an entire study population for a specific behavior, which is very unlikely to occur (and naturally warrants the question how useful a behavior would be for analysis).

Because our data indeed was compiled from studies that compared behavior between different inbred strains, we used to the pooled data (the combined data across trials of all experimental groups in all included studies) as a reference group to normalize our variables to z-scores. For each behavioral measure from appendix Table A2, z-scores for individual animals were calculated using the formula below, which indicates how many standard deviation (SD, σ) an observation (X) is above or below the mean (μ) of the pooled data:

$$z = \frac{X - \mu}{\sigma}$$

Although it is not common to treat discrete numerical data as continuous, the means and SD for *'total number-variables'* were also calculated and subsequently z-scores were computed; i.e. these so-called count based measures were treated as continuous data as suggested by Fagerland et al. (2011). The computed z-scores for single behavioral mHB measures were subsequently averaged within each behavioral dimension. In this procedure, the directionality of z-scores was adjusted so that increased score values reflected increased values for that behavioral dimension. This is illustrated in the example below for the behavioural dimension 'Risk assessment', which included the variables 'total number of stretched attends', and 'latency to the first stretched attend'.

T = total number of stretched attends; ***L*** = latency until first stretched attend; ***R*** = risk assessment

$$\mathrm{zr_T} = \frac{\mathrm{X_T} - \mu_\mathrm{T}}{\sigma\mathrm{T}};\ \ \mathrm{z_L} = \frac{\mathrm{X_L} - \mu_\mathrm{L}}{\sigma\mathrm{L}};\ \ \mathrm{Z_R} = \frac{\mathrm{Z_T} + (-\mathrm{Z_L})}{2}$$

2.6. Statistical analyses

All analyses were conducted with R version 3.5.1 in R-Studio (R Core Team, 2018). All Figures were created with GraphPad Prism (GraphPad Prism version 7.04 for Windows, GraphPad Software, La Jolla, California USA, www.graphpad.com).

2.6.1. Residuals for clustering: linear mixed models

The procedure described in Section 2.5 yielded five trajectories of integrated behavioral z-scores for each individual mouse, one trajectory per behavioral dimension. These five trajectories were subsequently fit with generalized linear mixed models to control for potentially confounding factors. The resulting standardized Pearson residuals could then be used for a clustering procedure.

Most of the potentially confounding factors were recoded into a single categorical variable. As listed in appendix Table A1, the included studies differed with respect to test location [3], experimenter [4], (sub-)strain [6], age [2], light condition [2] and sex [2] (number of categories in brackets). The majority of these factors consisted of only a few levels, causing risk for collinearity. We therefore summarized them in the categorical variable 'Group', yielding 13 levels. Other included explanatory variables were day of test to control for seasonal effects, counting from the first day of the year a particular trial was run (Ferguson and Maier, 2013) and test order (within a single test day) to control for time of day effects (Chesler et al., 2002). The variable 'trial' was intentionally left out of the model because we wanted to maintain this information in the residuals so that we could assess behavioral responses of individual mice over time (trials).

Linear mixed effects models were run using the package 'nlme' (Pinheiro et al., 2018). All models included Group and day of test as fixed predictors (without interaction). Individual intercepts (mouse ID) as well as intercepts for time of day, nested within individual mice, were included as random factors. Model assumptions were assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots (Sokal and Rohlf, 1995; Zuur et al., 2009). The variable arousal was logarithmically transformed to achieve normality of the residuals. Heteroscedasticity was avoided using the 'varIdent' variance structure transformation from the 'nlme' package when needed. This particular transformation allowed different residual spread for each level of the categorical variable 'Group' in our model (Zuur et al., 2009), and was applied on all five dimensions.

2.6.2. Cluster analysis

The resulting standardized Pearson residual z-score trajectories were subsequently analyzed with a k-means clustering procedure using the package 'kml3d' (Genolini et al., 2015). The advantage of the 'kml3d' procedure is twofold: it allows for dependence between time points (as is nearly always the case in longitudinal data) and it allows for analysis of joint response trajectories: multiple continuous response variables that were collected on the same instance (Genolini et al., 2015). Our joint response trajectories consisted of the five behavioral trajectories for each individual mouse. These were clustered simultaneously to explore the occurrence of homogeneous groups of mice that follow the same response on all five behavioral dimensions.

Prior to analysis the gap statistic was applied to evaluate whether the data was perhaps best represented by a single cluster, using the package 'clusGap' (Tibshirani et al., 2002). 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 (Tibshirani et al., 2002), and as such gives an indication of whether it is appropriate to partition the data into clusters. Using the kml3d-algorithm, partitioning into $k = 2$ to $k = 6$ clusters was assessed with the 'nearlyAll' configuration, using Euclidean distance as distance measure and Copy Mean for monotone missing values for imputation of missing values (see Genolini et al., 2015 for a detailed description of these settings). The analysis compiled 1000 iterations for each k clusters between 2 and 6, resulting in 5000 cluster solutions.

2.6.3. Cluster selection

The optimal partitioning of the clusters was selected using the approach of Clustering Validity Indices (CVI's) as described by Kryszczuk and Hurley (2010). CVI's combine indices from multiple quality criteria and as such form an effective strategy to optimize accuracy in cluster number selection (Wahl et al., 2014). The selection criteria that were used were Calinski-Harabasz, Ray-Turi and Davies-Bouldin (Genolini et al., 2015). These three non-parametric criteria reflect the relative compactness within clusters versus distance between clusters (Genolini and Falissard, 2010). The higher the value for Calinski-Harabasz, the more compact the clusters and the larger the differences between clusters. Conversely, high values for Ray-Turi and Davies-Bouldin reflect less compactness within clusters and smaller distances between clusters. To make the three criteria comparable, we used negative values for Ray-Turi and Davies-Bouldin, and criteria were normalized to values between 0 and 1 according to the following formula (Wahl et al., 2014):

$$Z_{i=}\frac{x_i - \min(x)}{\max(x) - min(x)}$$

The optimal number of clusters ($k = 2$ to $k = 6$ clusters) was selected according to the procedure suggested by Wahl et al. (2014). First, the optimal partition according to the Calinski-Harabasz criterium was selected for each scenario of $k = 2$ to $k = 6$ clusters. The arithmetic mean of the three quality criteria (the fused CVI) on that partition was then computed for each number of clusters. The cluster number with the highest fused CVI was subsequently selected as the optimal cluster number.

2.6.4. Cluster characterization

The obtained clusters were characterized by linear mixed models that analyzed the difference between clusters in residual integrated behavioral z-scores over trials. The main model for each behavioral dimension included cluster, trial and their interaction as fixed predictors. Individual intercept (mouse ID) was included as random factor. Individual slope (trial nested in mouse ID) was initially also included as random factor, but was ultimately left out of the models as the correlation between individual slopes and intercepts was near perfect (r < -0.992 in all models), which may reflect overparameterization and result in loss of power (Matuschek et al. 2017). Models were run with a continuous autoregressive correlation structure (AR(1) process for a continuous time covariate) and fit with restricted maximum likelihood.

Model assumptions were again assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots. A square root transformation was applied on the residual integrated z-score for risk assessment to achieve normality of the residuals. Heteroscedasticity was avoided using the 'varIdent' variance structure transformation from the 'nlme' package when needed. The models for the variables avoidance behavior, risk assessment and exploration included a transformation that allowed differential residual spread between clusters. The model for locomotion included a transformation that allowed differential residual spread between trials.

Significant main and/or interaction effects were further broken down by *post hoc* tests using the package 'emmeans', which enables users to obtain least squares means for linear mixed models and compute contrasts for *post hoc* assessment (Lenth, 2019).

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. The α was computed using the following formula: $\alpha = 1\text{-}[1 - 0.05]^{1/\lambda}$, where $\lambda =$ the number of times a group was used in a comparison. For all five behavioral dimensions, general directionality of the response curve for each cluster was assessed by pairwise comparisons of the estimated marginal means between trial 1 and trial 20 (the first and the last trial of testing, $\alpha = 0.02532$). In addition, differences in onset levels of behavior between clusters were assessed by *post hoc* comparisons of the estimated marginal means on trial 1 ($\alpha = 0.05$). For the behavioral dimensions risk assessment and arousal additional *post hoc* tests were conducted to assess the differences in (estimated marginal means) between clusters on each trial. For these specific comparisons the α was set to 0.00256, again using the Dunn-Šidák correction.

Main and interaction effects from the linear mixed models were derived using conditional F-tests with corresponding P value ($\alpha = 0.05$). All *post hoc* contrasts were summarized as the difference between the two estimated marginal means and their corresponding standard error, t statistic, and P values. In addition, Cohen's d effect size was reported to estimate the relative weight of *post hoc* comparisons. Cohen's d was computed from the value of the t test that resulted from the pairwise comparisons, with the following formula, where t represents the value of the t test between two clusters, and n1 and n2 the respective sizes of each cluster (Rosenthal and Rosnow, 2008):

$$\frac{t(n1 + n2)}{sqrt(df)^{*}\, sqrt(n1n2)}$$

The guidelines provided by Wahlsten (2011) 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, $|d| > 1.5$.

Residual integrated behavioral z-scores for clusters on each dimension were summarized as means with 95 % confidence intervals in Fig. 1. The differences between clusters in residual integrated behavioral z-scores on trial 1 were graphed as means with 95 % confidence intervals in Fig. 2.

2.6.5. Cluster stability

Stability of the clusters was assessed by a bootstrapping procedure in which 200 random samples (of n = 125) 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 (Clatworthy et al., 2005). 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 (Fig. 3).

2.7. Ethical note

All experimental protocols were approved by the Animal Experiments Committee of the Academic Biomedical Center Utrecht,

Table 1
Cluster size and distribution of (sub-) strains across clusters.

Cluster size (n) and proportion of total n per cluster				
	Cluster A		Cluster B	
n total = 125	n = 73 *(58.4%)*		n = 52 *(41.6%)*	
Distribution of strains *within* clusters				
	Cluster A		Cluster B	
(sub-) Strain	n	%	n	%
BALBc/J	–	–	40	76.9
129P3/J	41	56.2	12	23.1
129P2/OlaHsd	8	10.9	–	–
129X1/J	8	10.9	–	–
129S2/SvPasCrl	8	10.9	–	–
129S2/SvHsd	8	10.9	–	–

Top row: Cluster size (n) and proportion of total population per cluster. Bottom rows: Distribution of (sub-) strains (n and proportion per strain) within each cluster.

the Netherlands (for approval numbers see supplementary Table A1). Decision for approval was based on the Dutch implementation of the EC Directive 86/609/EEC (Directive for the Protection of Vertebrate Animals Used for Experimental and Other Scientific Purposes; Anonymous, 1986). Furthermore, the experiments followed '*the Principles of Laboratory Animal Care*' and refer to the '*Guidelines for the Care and Use of Mammals in Neuroscience and Behavioral Research*' (National Research Council, 2003). Finally, all experiments were reported in accordance with the ARRIVE-guidelines to the author's best ability (http://www.nc3rs.org.uk/arrive-guidelines; Kilkenny et al., 2010).

3. Results

3.1. Cluster analysis

The optimal partition of the data yielded two clusters. Selection of the optimal partition was based on the CVI of three quality criteria: Calinski-Harabasz, Ray-Turi and Davies-Bouldin (Results not shown). Cluster size and distribution of (sub-) strains across clusters were presented in Table 1. The majority of mice grouped together in cluster A (58.4 %). This cluster was composed of the majority of 129P3/J mice (77.4 %) and all mice of the four other 129 sub-strains. The remaining 129P3/J mice formed cluster B, together with all BALB/cJ individuals.

3.2. Cluster characterization

To characterize the clusters on each behavioral dimension, linear mixed models were conducted to analyze between-cluster differences across trials. These results are presented for each dimension in subheadings *3.2.1 – 3.2.5*. Section *3.2.6* provides a summary description of the different response types in each cluster. A visual representation of the behavioral response across trials in clusters A and B for each dimension is depicted in Fig. 2 (presented as mean residual integrated behavioral z-scores with 95 % CI). Fig. 3 shows the mean levels of behavior on the first trial for each cluster, again in each dimension (summarized as estimated marginal means with 95 % CI).

3.2.1. (Residual integrated z-score for) Avoidance behavior

Avoidance behavior was predicted by trial ($F_{(19,\ 2373)} = 3.51$, $P < 0.0001$), but this effect was confounded by a significant interaction between cluster and trial ($F_{(19,\ 2373)} = 47.54$, $P < 0.0001$), see Fig. 2. Pairwise comparisons of the estimated marginal means between trial 1 and trial 20 were conducted separately for each cluster to characterize the directionality of avoidance slopes. Mice in cluster A displayed a significant increase in avoidance behavior (-2.020 ± 0.128, $t_{(2337)}$ = -15.723, $P < 0.0001$), while mice in cluster B significantly decreased avoidance behavior between the first and the last trial (2.073 ± 0.144, $t_{(2337)}$ = 14.364, $P < 0.0001$), both with moderate effect sizes (d = -0.650 and d = 0.594 respectively).

In addition to differences in the course of avoidance behavior over trials, we assessed cluster differences in onset levels of avoidance behavior. *Post hoc* comparisons of the estimated marginal means revealed statistical differences on trial 1 between clusters A and B (-3.043 ± 0.137, $t_{(123)}$ = -22.275, $P < 0.0001$) with a very large effect size (d = -4.075), see Fig. 3.

The significant interaction between trial and cluster could thus be explained by the contrasting patterns in avoidance behavior between the clusters: mice in cluster A increased avoidance behavior while cluster B decreased avoidance behavior as trials progressed.

3.2.2. (Residual integrated z-score for) Risk Assessment

Risk assessment was significantly predicted by trial ($F_{(19,\ 2335)} = 94.64$, $P < 0.0001$), but this effect was confounded by a significant interaction between cluster and trial ($F_{(19,\ 2335)} = 4.45$, $P = 0.0001$), see Fig. 2. *Post hoc* comparisons of the estimated marginal means between trial 1 and trial 20 indicated that in both cluster A (0.561 ± 0.038, $t_{(2335)}$ = 14.807, $P < 0.0001$, d = 0.613) and cluster B (0.793 ± 0.031, $t_{(2335)}$ = 25.698, $P < 0.0001$, d = 1.064) risk assessment decreased significantly between the first and the last trial, with medium and large effect sizes respectively.

However, pairwise comparisons of the estimated marginal means between clusters on each of the 20 trials (adjusted $\alpha = 0.00256$) showed that clusters only differed in risk assessment on trial 1 (-0.217 ± 0.035, $t_{(123)}$ = -6.272, $P < 0.0001$, see Fig. 3) and trial 2 (-0.190 ± 0.034, $t_{(123)}$ = -5.509, $P < 0.0001$), with a large effect size for trial 1 (d = -1.131), and a moderate effect size for trial 2 (d = -0.993). The significant interaction between cluster and trial thus appeared to be predominantly driven by an effect of trial (a general decrease in risk assessment), and the fact mice in cluster B displayed higher onset levels of risk assessment.

3.2.3. (Residual integrated z-score for) Arousal

The main model indicated a significant effect of trial ($F_{(19,\ 2337)} = 6.24$, $P < 0.0001$), but this effect was confounded by a significant interaction between cluster and trial ($F_{(19,\ 2337)} = 2.40$, $P = 0.0006$), see Fig. 2. Visual inspection of the data (Fig. 2) however, suggested that arousal curves were highly similar between clusters. *Post hoc* tests comparing the estimated means between trials 1 and 20 indicated that neither cluster displayed a significant change in arousal across trials (A, -0.262 ± 0.157, $t_{(2337)}$ = -1.672, P = 0.0946, d = -0.069; B, -0.371 ± 0.186, $t_{(2337)}$ = -2.000, P = 0.0456, d = -0.082).

The significant interaction between trial and cluster was thus further explored by pairwise comparisons of the estimated marginal means between clusters on each trial (adjusted $\alpha = 0.00256$). This revealed that clusters only differed in estimated means of arousal on trial 4 (0.758 ± 0.172, $t_{(123)}$ = 4.411, $P < 0.0001$), with a moderate effect size (d = 0.795). It was therefore concluded that the significant effects in the main model may have been the result of minimal fluctuation in arousal across trials in combination with potential over-parametrization of the model, rather than the reflection of meaningful differences between clusters.

3.2.4. (Residual integrated z-score for) Exploration

Exploration was significantly predicted by trial ($F_{(19,\ 2335)} = 11.80$, $P < 0.0001$) but this effect was confounded by a significant interaction between cluster and trial ($F_{(19,\ 2335)} = 29.96$, $P < 0.0001$), see Fig. 2. *Post hoc* comparisons of the estimated marginal means between trials 1 and 20 showed that cluster A displayed a significant decrease in exploration with a small effect size (0.839 ± 0.150, $t_{(2335)}$ = 5.577, $P < 0.0001$, d = 0.231) while cluster B significantly increased exploration as trials progressed (-2.216 ± 0.141, $t_{(2335)}$ = -15.699, $P < 0.0001$) with a moderate effect size (d = -0.650). Onset levels of exploration were higher for cluster A than for cluster B, with a very large effect size, as indicated by a *post hoc* test comparing mean

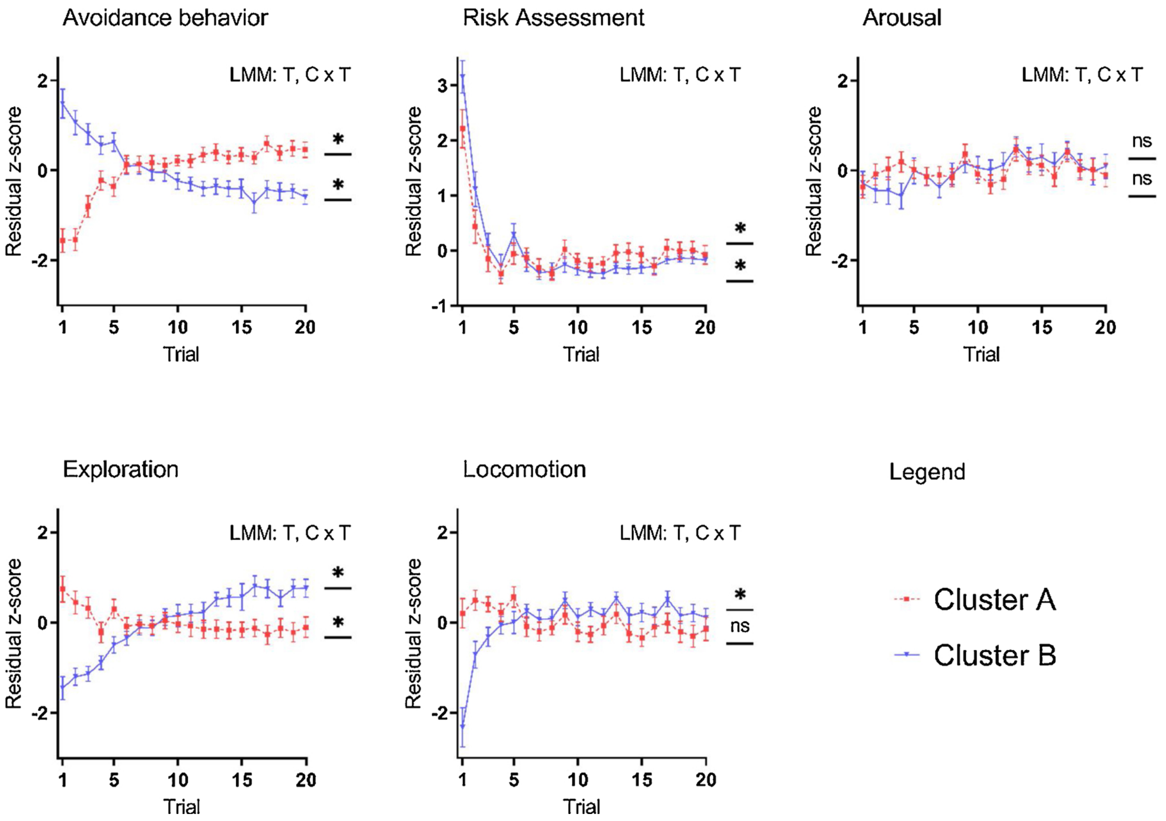

Fig. 2. Residual integrated behavioral z-scores for mice in clusters A and B.
Results are presented as means with 95 % CI. Effects were significant in the linear mixed models (LMM) when $P < 0.05$. T indicates a significant main effect of Trial; CxT indicates a significant interaction between cluster and trial. * = Significant ($P < 0.02532$) *post hoc* comparison of the estimated marginal means between trial 1 and trial 20 for each cluster. ns = non-significant difference in *post hoc* comparison between trial 1 and trial 20. Note: Risk assessment scale on the y-axis differs from the other four dimensions.

exploration on trial 1 (2.194 ± 0.146, $t_{(123)} = 15.039$ $P < 0.0001$, $d = 2.751$), see Fig. 3.

These between cluster differences in onset levels and contrasting curves of exploration across trials underlie the general interaction between trial and cluster: mice in cluster B increased exploratory behavior as trials progressed while mice in cluster A decreased this type of behavior.

3.2.5. *Residual integrated z-score for) locomotion*

Locomotion was predicted by a significant effect of trial ($F_{(19, 2336)} = 7.52$, $P < 0.0001$) and a significant interaction between cluster and trial ($F_{(19, 2336)} = 10.64$, $P < 0.0001$), see Fig. 2. *Post hoc* comparisons of the estimated marginal means for each cluster between trial 1 and trial 20 showed that mice in cluster A did not display a change in locomotion (0.351 ± 0.195, $t_{(2336)} = 1.794$, $P = 0.0729$),

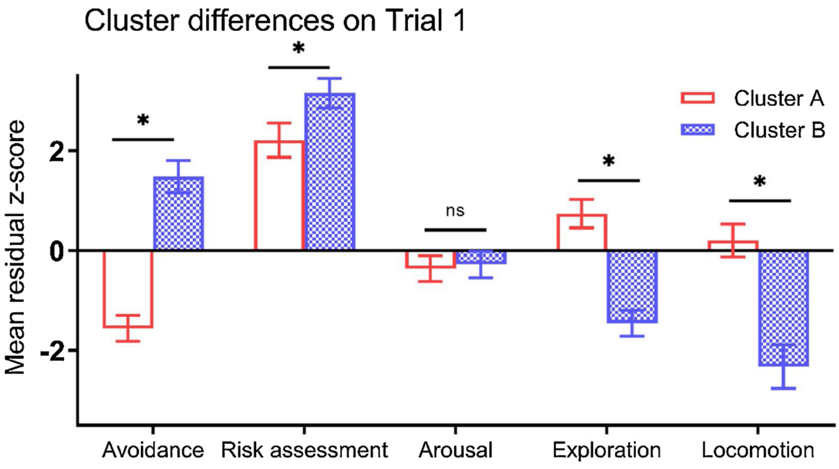

Fig. 3. Initial levels on al behavioral dimensions for clusters A and B (mean residual z-score on trial 1). Results are summarized as estimated marginal means with 95 % confidence intervals. * = Significant difference ($P < 0.05$) in *post hoc* comparison, ns = difference not significant.

while mice in cluster B increased locomotion between the first and the last trial (-2.432 ± 0.233, $t_{(2336)}$ = -10.438, $P < 0.0001$), both with small effect sizes (respectively $d = 0.074$ and d = -0.432). *Post hoc* comparisons of mean locomotion on trial 1 furthermore showed that clusters differed in initial levels of locomotion (2.526 ± 0.253, $t_{(123)}$ = 9.993, $P < 0.0001$), with a very large effect size ($d = 1.827$), Fig. 3. Thus, the significant interaction between cluster and trial appears predominantly driven by the fact that mice in cluster B increased locomotion, while mice in cluster A did not change locomotor activity as trials progressed.

3.2.6. Summary characterization of clusters

The clusters were characterized by significantly contrasting patterns in anxiety related behavior and activity patterns. Most notably, mice in cluster A increased avoidance behavior, while avoidance decreased in cluster B after repeated exposure to the test. In rodents, behavior displayed in a novel environment is often regarded as the net result of conflict between the motivation to avoid a potentially harmful situation and the drive to explore the novel stimulus (the approach/avoidance conflict). In cluster B, a decrease in avoidance behavior was coupled with an increase in exploration and locomotion. Initial inhibition of the drive to explore was lifted once the situation was assessed to be safe, resulting in habituation. The profile of cluster B was highly similar to BALB/cJ response that was observed in original studies, which was classified as habituating to the test. This cluster can thus be characterized as ‘habituation profile’.

Mice in cluster A however, increased avoidance behavior, while exploration decreased and locomotion remained unchanged across trials. This profile was reminiscent of the sensitization response that was observed in 129-mice in the original data, which reflected failure to habituate to the test. This cluster can thus be classified as a ‘sensitization profile’. The profile of cluster A also differed in an important aspect. In the original studies, sensitization was predominantly indicated by an increase in avoidance behavior, but changes in exploration and locomotion were less pronounced, while one would expect a decrease in activity patterns according to the approach/avoidance conflict described above. This (decrease) is indeed what was found for exploratory behavior in cluster A. Zooming in on individual responses of 129-mice thus revealed a more pronounced sensitization profile compared to the original studies. Finally, risk assessment and arousal did not differ between the clusters.

3.3. Cluster stability

The 200 clustering solutions from the bootstrap samples appear highly comparable to the original solution. Fig. 4 depicts the mean trajectory of all 200 samples (black dashed line) against the trajectory belonging to the original cluster (red, cluster A; blue, cluster B), as well as the trajectories of each bootstrap sample (grey) for each cluster, on each dimension. The average Jaccard similarity index for cluster A was 0.96, meaning that on average, an individual mouse belonged to cluster A in 96 % of the bootstrap samples. The average Jaccard similarity index for cluster B was 0.93. Fig. 4 depicts all individual Jaccard similarity indices per cluster, their associated mean and sd. All in all, these results indicate that the identified clusters are stable.

3.4. Relative weight of dimensions on clustering solution

The clusters described above were partitioned on all five behavioral dimensions. However, differential cluster responses were more pronounced on some dimensions than on others, with significant differences between clusters in avoidance behavior, exploration and locomotion, but largely similar patterns of arousal and risk assessment. Therefore, we wanted to test whether some dimensions were perhaps more ‘influential’ in the partitioning of response types than others. We conducted an additional series of cluster analyses, all with four dimensions, each time leaving one of the five dimensions out. Pearson Chi Square tests were used to assess whether the cluster size in these analyses deviated from the partitioning that was obtained with five dimensions. In addition to this, the number of individual mice that fell into a different cluster after excluding a certain behavioral dimension was recorded. Table 2 gives an overview of the cluster sizes for each of the analyses. Although excluding a single dimension from the cluster analysis did result in slight changes in cluster size and composition for some dimensions, none of these changes were significantly different from the original partitioning.

In an increasing order with respect to impact: Omitting arousal yielded the exact same clusters ($X^2_{(1)}$ = 0.000, P = 1.000), with a distribution of mice across clusters that was identical to the distribution based on five dimensions (none of the mice fell in a different cluster). After excluding risk assessment, one single mouse fell in another cluster and cluster sizes were highly similar to the results based on five dimensions ($X^2_{(1)}$ = 0.017, P = 0.898), see Table 2. In the case of locomotion, five individuals ‘switched’ cluster, but cluster sizes were not significantly different from the distribution based on five dimensions ($X^2_{(1)}$ = 0.150, P = 0.699). Omitting avoidance or exploration resulted in the most substantial change in cluster size and distribution of individuals: in both analyses, eight individuals fell in a different cluster compared to the distribution based on five dimensions, but changes in cluster size were not significant for avoidance behavior ($X^2_{(1)}$ = 0.261, P = 0.610) or exploration ($X^2_{(1)}$ = 1.082, P = 0.298). These results suggest that although none of the five dimensions dominated the partitioning of the clusters, some were more influential than others. Exploration and avoidance behavior exerted the most weight on partitioning of the response types, while the contribution of arousal and risk assessment was relatively small.

4. Discussion

The current paper explored inter-individual variability in habituation and sensitization responses in two mouse inbred strains. We reinspected data from a series of studies that measured impaired habituation to a novel environment as a possible indicator for non-adaptive, i.e. pathological anxiety in BALB/cJ and various 129-substrains (Salomons et al., 2010a, b; Salomons et al., 2010c, 2013; Boleij et al., 2012).

In these mechanisms, the temporal progression of a response is essential for assessing its adaptive quality. Also, anxiety related behavior is typically expressed by a combination of behavioral dimensions (Rodgers and Dalvi, 1997; Belzung and Griebel, 2001; Ohl, 2003; O’Leary et al., 2013). Our objective therefore was to take each individual response trajectory into account in analysis, and assess whether clustering these individual trajectories would identify subgroups of response that grouped together across multiple behavioral dimensions. This resulted in two homogenous subgroups of mice, representing a habituation and a sensitization response profile.

4.1. Benefits

Overall, the habituation and sensitization profiles that emerged from these analyses mirrored the two contrasting phenotypes from that were identified by comparing average strain responses in the original

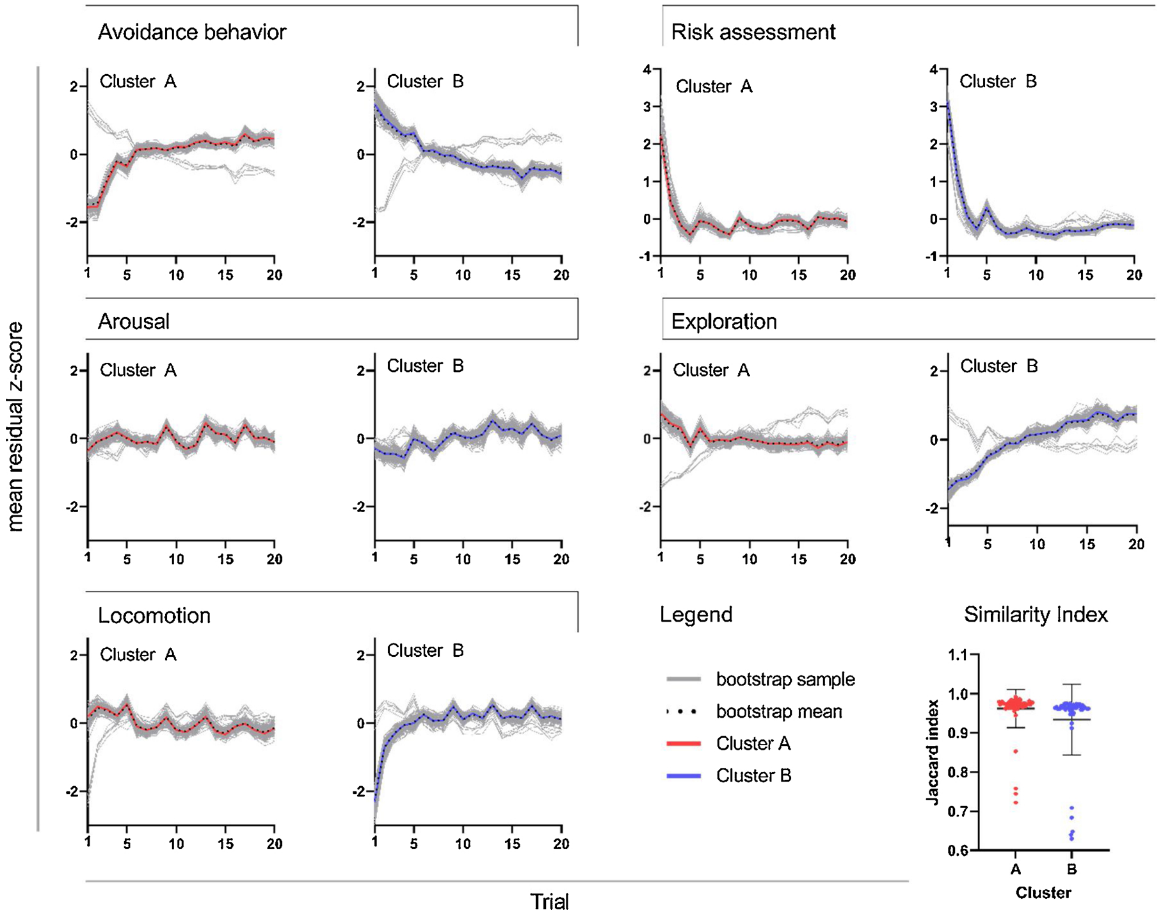

Fig. 4. Mean trajectories (residual z-scores) of 200 bootstrap samples (grey) for each cluster, on each behavioral dimension. For visual comparison, the average trajectory of all bootstrap samples is depicted (black dotted line) against the trajectory belonging to cluster A (red) or B (blue). These two trajectories are highly similar across behavioral dimensions. Last panel: Jaccard similarity index for clusters A and B (individual points, and population mean and sd).

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

Cluster	All dimensions	Excluded				
		AVO[a]	RA[a]	AR[a]	EXPL[a]	LOC[a]
	N	n	n	n	n	n
A	73	69	74	73	81	76
B	52	56	51	52	44	49
Total	125	125	125	125	125	125

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

studies. Interestingly however, our analyses also yielded new information.

First, it demonstrated that subtypes of response may occur within the same inbred strain. 129 mice were found to display both habituation and sensitization profiles when exposed to a novel environment, while BALB/cJ mice showed less within strain variation by consistently displaying a habituation response. The prevalence of subtypes of emotional response within the same inbred strain is not new. Within strain variation in anxiety responses has been previously documented in BALB/cJ and 129 J mice (Ducottet and Belzung, 2004; Cohen et al. 2008; Jakovcevski et al. 2008). Our results are partially consistent with these findings, although we did not observe within strain variability in BALB/cJ. Mouse inbred strains may however also differ in their phenotypic robustness, resulting in differences in within strain variability between strains. In an extensive study comparing within strain variability in 8 isogenic strains, Loos et al. (2015) demonstrated that BALB/cJ mice ranked low in within strain variability while at the same time, 129S1/Sv mice (not included here) showed reduced phenotypic robustness, leading to high within strain variability (Loos et al., 2015). Our results suggest that this may also pertain to other 129-substrains but the relatively small number of included datasets in our analyses limits the possibility of drawing vast conclusions.

Second, the analyses showed that individual mice consistently grouped together on multiple behavioral dimensions. This is in line with other findings that anxiety related behaviors (such as behavioral habituation) are expressed by multiple behavioral dimensions (Belzung and Griebel, 2001; Ohl, 2003; O'Leary et al., 2013; Labots et al., 2016). It also indirectly seems to support the notion that behavioral habituation and sensitization in rodents is a complex phenomenon that involves sensory, cognitive and emotional processes (Bolivar, 2009; Boleij et al., 2012).

Third, zooming in on individual response curves yielded a more pronounced sensitization response than was initially observed in the original data. In our analyses, the sensitization profile was characterized by an increase in avoidance and a decrease in exploration, while in the original studies, sensitization was primarily indicated by an increase in avoidance behavior only and changes in activity parameters were less pronounced. These three findings illustrate that an individual based approach may complement analyses based on group effects.

As noted before, more detailed information about individual variability and subtypes of response within the data may contribute to the quality of animal experiments. Lonsdorf and Merz (2017) for example argued that the existence of subpopulations within a study sample displaying contrasting response patterns may mask the detection of significant differences on group level (i.e. a type II error).

Also, the identification of subgroups of individuals that show the same response pattern may also prove of valuable interest within the context of systematic heterogenization (Bodden et al., 2019). This concept was advocated by Richter (2017) and entails the systematic introduction of factors that affect variation in observed results as a means to increase robustness of experimental findings. Although simulation studies have indicated a promising effect on increasing reliability of results, the challenge remains to identify factors that affect variation which are suitable (and workable) for systematic variation within a single experiment (Richter, 2017; Bodden et al., 2019). Potential factors that have been suggested are batch, experimenter and testing time (Paylor, 2009; Richter, 2017; Bodden et al., 2019. Systematic variation of individual response profiles may prove another suitable factor that could be varied across experimental groups.

From a translational perspective, the identification of sub-profiles of anxiety related behavior in mouse models may help to gain better insight into the underlying mechanisms responsible for differential vulnerability for anxiety disorders in humans (Einat et al., 2018; Stegman et al. 2019). In a clinical situation, exposure to a similar condition may result in development of affective disorders in some, while other people are unaffected. Kazavchinsky et al. (2019) therefore argue that corresponding animal models should attempt to explore similar patterns of responding.

The multivariate, longitudinal based, clustering approach utilized in this paper may also be of interested in other domains. The integration of multiple measures as a means to assess individuality has not only been advocated for emotional reactivity (Ramos and Mormède, 1998; Hager et al., 2014), but also for other constructs such as coping style (Koolhaas et al., 2010; Koolhaas and van Reenen, 2016), behavioral syndromes (Bell, 2007) and temperament (Réale et al., 2007; Finkemeijer et al., 2018). Koolhaas and van Reenen (2016) for example proposed a 3-dimensionsal model using coping style, emotionality and sociality to assess individual vulnerability to stress related diseases. Similarly, Reále et al. (2007) emphasized the combined analysis of traits to describe the full nature of temperament.

Multidimensionality is typically assessed by multivariate approaches such as principal components analysis (PCA) or factor analysis. When one studies traits that heavily rely on the temporal/longitudinal nature of response however, these approaches offer no avail. In that light, the kml3d-clustering algorithm employed in the present study constitutes a valuable addition to the available techniques.

4.2. Limitations

While the benefits of taking individual variation into account are evident, the applied clustering approach from this paper also has its drawbacks. The first limitation is inherent to clustering techniques in general. These techniques are mainly exploratory and do not statistically infer the reality of existence of the clusters (Genolini et al., 2015). In other words, there is no single reliable method to determine the “true” number of clusters in a given dataset (Genolini et al., 2015; Everitt et al. 2001). In our analyses we had no a priori assumptions regarding the number of clusters as this was the first time the kml3d-clustering approach was applied to habituation and sensitization responses. Therefore we used the method proposed by Kryszczuk and Hurley (2010), and adjusted by Wahl et al. (2014), that combined three commonly applied quality criteria to a single clustering validity index (CVI) as a means to select the optimal partition. This method has been proven a validated way to increase robustness and accuracy of cluster number selection in comparison to a single quality criterion (Kryszczuk and Hurley, 2010).

Secondly, although no single dimension was dominant in partitioning of the clusters, some dimensions appeared more influential in determination of response types than others (Table 2). Contrasts between clusters were most evident in avoidance behavior and exploration, which can be interpreted by the interplay between avoidance and exploratory behavior (the approach/avoidance conflict, Ohl, 2003). Exploration is inhibited by anxiety, and as such represents an indirect measure of anxiety (Ohl, 2003). When taking previous studies in the mHB into account, it seems hardly surprising that these dimensions constituted the most defining factors in partitioning of our clusters. Avoidance behavior was the most distinguishing feature between habituation and sensitization in the original studies (Salomons et al., 2010a, b; Salomons et al., 2010c, 2013; Boleij et al., 2012). Also, behaviors indicative of avoidance behavior and exploration formed the two largest components in a principal component analysis (PCA) summarizing behaviors measured in the mHB (explaining 36.7 % of the total variance, Laarakker et al., 2008). In a later study by Labots et al. (2016) the same mHB-based composite z-scores that were used in the present analyses were found to correlate strongly with components that were obtained in the PCA by Laarakker et al. (2008).

The impact of locomotion on partitioning of the clusters was lower than for avoidance behavior and exploration. Like exploration, locomotor activity is not only associated with general activity levels but also has a confounding effect on anxiety related behavior (O’Leary et al., 2013). In fact, an alternative interpretation of anxiety related behavior is that differences in (lack of) exploration of a specific area may just as well be the result of differences in overall activity levels (Boleij et al., 2012). In the context of anxiety studies, it is therefore important to distinguish between horizontal (e.g. line crossings in the mHB) and vertical activity (e.g. rearing behavior). In the present study this distinction was indeed included, with rearing behavior regarded an exploratory activity, while the dimension locomotion only included horizontal activity.

Locomotor activity is also strongly strain-specific (O’Leary et al., 2013) and some 129 strains are indeed known for their low levels of locomotion and exploratory activity (Cook et al., 2002; Boleij et al., 2012). A persisting high level of avoidance behavior was indeed combined with low locomotor and explorative activity in two 129S2 strains (Boleij et al., 2012), but not in the remainder (and majority) of the included 129 strains. It thus seems unlikely that a potential confounding effect of locomotion was the reason for its lower impact in the partitioning of the clusters.

Perhaps this lower impact may be better explained by the fact that differences in locomotion between clusters were less pronounced over trials (compared to avoidance behavior and exploration). Although the analyses indicated that clusters differed significantly in locomotion (by means of a significant interaction between cluster and trial), *post hoc* analyses revealed that this effect was predominantly driven by locomotion differences on the first 5 trials. On the remaining trials, locomotion was largely similar between clusters.

A similar explanation may account for the relatively low impact of risk assessment. Inferential analysis of the clusters indicated that clusters differed in display of risk assessment across trials, but *post hoc* inspection revealed that clusters only differed in initial levels of this behavior: mice in cluster A showed lower levels of risk assessment in the first two trials than mice in cluster B. Finally, the absence of discriminative weight for arousal was hardly surprising as neither cluster displayed a change in arousal across trials and clusters did not significantly differ between one another.

All in all, this illustrates a potential pitfall of the utilized clustering approach. The fact that risk assessment, arousal and locomotion exerted a smaller discriminative effect could also imply that more subtle effects are conflated when pooling all scored behaviors in a single analysis. Genolini et al. (2015) addressed this point by stating that the relative weight of variables can be of issue when partitioning joint trajectories. They relate this matter however to variables that are measured on different scales, and provide a function to standardize the variables in their algorithm to overcome this issue. As our data was already summarized in composite z-scores this could not have been the issue here. If anything, this indicates that one should be considerate of which variables/dimensions are included when clustering joint trajectories. When the goal is to identify individuality in behaviors that are expressed in more low frequencies or which are very strain/species dependent perhaps a univariate cluster analysis is more desirable.

Also, a relatively small portion of the mice was female (BALB/cJ, n = 10; 129P3/J, n = 10; Salomons et al., 2010a). Female mice are traditionally underrepresented in preclinical research, mainly because of the assumption that that females show more variability in response due to their estrous cycle (Mogil and Chanda, 2005; Prendergast et al., 2014). In an extensive review comparing variability between male and female mice however, Prendergast et al. (2014) found that females are no more variable than males. In the same fashion, several studies found that females tested at random points in their estrous cycle do not differ in variability from males (Mogil and Chanda, 2005; Laarakker et al., 2011). The present individual-based analyses extend these results, although the small sample size makes it difficult to draw vast conclusions. BALB/cJ females all displayed a habituation response, in agreement with all BALB/cJ males. Females of the 129P3/J showed even less variability in response than males, as all females displayed a sensitization response (cluster A) while 22.6 % (n = 12) of the 129P3/J males deviated from the response that was displayed by the majority of 129-mice and grouped together in cluster B.

Incorporating sex as a discerning factor in rodent models of psychopathologies has become increasingly advocated in the last decade (Kokras and Dalla, 2014; Prendergast et al., 2014). Incorporating female findings preclinical anxiety research is especially relevant as anxiety disorders are more prevalent in women then in men (Zender and Olshansky, 2009) and factors such as clinical course and treatment response are known to differ between sexes (Donner and Lowry, 2013). To our knowledge however, only a few studies (e.g. Pitychoutis et al., 2011; Carreira et al., 2017; Kazavchinsky et al., 2019) have directly addressed sex differences in individual variability in rodent models. Further assessment of individual response profiles between and within sexes may provide additional insight to mechanisms underlying sexual dimorphism in vulnerability and response to treatment in human patients (Pitychoutis et al., 2011).

The last issue concerns the fact that the dataset used in these analyses was compiled of 7 different mHB-experiments (appendix Table A1). These studies were combined because clustering approaches require a substantial sample size to detect meaningful clusters (Dolnicar et al., 2016). These studies however, have been conducted over a time span of 4 years (2006–2010) and vary in factors that are known to affect variability between experiments, such as test location, experimenter, time of year etcetera (Crabbe et al., 1999; Garner, 2005). At this point it is unclear to what extent these factors accounted for (part of) the variation that resulted in the partitioning of the clusters. Although the bootstrapping procedure indicated that these clusters were stable (Fig. 4), we believe that further validation of the obtained results is necessary in order to assess whether the identified variation in response profiles is robust and exemplary for BALB/c and 129-mice in general. This variation should ideally be empirically addressed in a study that is specifically designed for such purpose (i.e. in a single experiment and with a sufficient sample size).

5. Conclusions

For now, the present paper showed that re-analyzing habituation and sensitization responses on an individual level yields distinct groups of individuals that group together on multiple behavioral dimensions. The combined analysis of multiple dimensions thus allows for a full description of differential profiles of emotional response types. It also yielded new, more detailed information on the characteristics of these response types, and allowed for the identification of individuals that may deviate from their strain specific response. In that respect, the approach of quantifying individual response trajectories and assessing the presence of groups of animals that show the same phenotype across behavioral dimensions presents an additional avenue to the GLMM-based approaches already available in the literature on capturing individual variation in analysis. To what extent the observed response types are robust, and whether taking these differences into account affects reliability of results remains to be tested.

Funding

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

CRediT authorship contribution statement

Marloes H. van der Goot: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Hetty Boleij:** Investigation, Data curation. **Jan van den Broek:** Methodology. **Amber R. Salomons:** Investigation, Data curation. **Saskia S. Arndt:** Conceptualization, Writing - review & editing. **Hein A. van Lith:** Supervision, Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgements

The authors are grateful to the two anonymous reviewers for their valuable comments on the manuscript, which has undoubtedly helped the authors to improve the article.

Appendix A

Table A1
Overview of factors that varied between experiments.

Study	(Sub) strains	Supplier[a]	Sex	Age at test (wks)	Housing location[b]	Experimenter	Cage size	Light on board	Observer version	Approval no
Salomons et al. (2010a)	BALB/cJ (N = 10) 129P3/J (N = 10)	J	F	~8	NVI	D	Eurostandard Type II	120 lx	5.0	2007.I.01.007
Salomons et al. (2010b)	129P3/J (N = 13)	J	M	~9	UU-P	C	Eurostandard Type II L	120 lx	4.0	2007.I.01.007
Salomons et al. (2010c)	I. BALB/cJ (N = 8) 129P3/J (N = 8)	J	M	~9	UU-P	A	Eurostandard Type II L	50 lx	4.0	2007.I.01.007
	II.BALB/cJ (N = 8) 129P3/J (N = 8)	J	M	~9	UU-P	A	Eurostandard Type II L	120 lx	4.0	2007.I.01.007
Salomons et al. (2013)	BALB/cJ (N = 14) 129P3/J (N = 14)	J	M	~8	NVI	A	Eurostandard Type II	120 lx	5.0	2009.I.06.044
Boleij et al. (2012)	I.129S2/SvPasCrl (N = 8) 129S2/SvHsd (N = 8)	Crl E*	M	~9	UU-P	A	Eurostandard Type II L	120 lx	5.0	2007.I.01.007
	II.129P2/OlaHsd (N = 8) 129X1/J (N = 8)	E*, J*	M	~9	UU-G	B	Eurostandard Type II L	120 lx	5.0	2009.I.10.079

[a] Supplier: J = Jackson Laboratory, Bar Harbor, ME, USA; Crl = Charles River Laboratories, ‘s-Hertogenbosch, the Netherlands; E = Envigo, Horst, the Netherlands*formerly Harlan Sprague Dawley inc., Horst, the Netherlands.

[b] Housing Location: NVI = National Vaccine Institute, Bilthoven, the Netherlands; 2 = Central Laboratory Animal Research Facility of Utrecht University, location Paviljoen, Utrecht, the Netherlands; 3 = Central Laboratory Animal Research Facility of Utrecht University, location GDL, Utrecht, the Netherlands.

Table A2
Behavioral variables measured in mHB and used for composition of z-scores in this paper.

Motivational system/Behavioral dimension	Behavioral variable	Directionality z-score[a]
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

[a] 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.

References

Allegue, H., Araya-Ajoy, Y.G., Dingemanse, N.J., Dochtermann, N.A., Garamszegi, L.Z., Nakagawa, S., Réale, D., Schielzeth, H., Westneat, D.F., Hadfield, J., 2017. Statistical quantification of individual differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. Methods Ecol. Evol. 8 (2), 257–267. https://doi.org/10.1111/2041-210X.12569.

Anonymous, 1986. Directive 86/609/EEC of 24 November 1986 on the Approximation of Laws, Regulations and Administrative Provisions of the Member States Regarding the Protection of Animals Used for Experimental and Other Scientific Purposes. OJEC L358:1-29.

Araya-Ajoy, Y.G., Mathot, K.J., Dingemanse, N.J., 2015. An approach to estimate short-term, long-term and reaction norm repeatability. Methods Ecol. Evol. 6 (12), 1462–1473. https://doi.org/10.1111/2041-210X.12430.

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

Bell, A.M., 2007. Future directions in behavioural syndromes research. Proc. R. Soc. B. 274 (1611), 755–761. https://doi.org/10.1098/rspb.2006.0199.

Bello, N.M., Renter, D.G., 2018. Reproducible research from noisy data: revisiting key statistical principles for the animal sciences. J. Dairy Sci. 101 (7), 5679–5701.

https://doi.org/10.3168/jds.2017-13978.

Belzung, C., Griebel, G., 2001. Measuring normal and pathological anxiety-like behavior in mice: a review. Behav. Brain Res. 125 (1-2), 141–149. https://doi.org/10.1016/S0166-4328(01)00291-1.

Beynen, A.C., 1991. The basis for standardization of animal experimentation. Scand. J. Lab. Anim. Sci. 18 (3), 95–99.

Beynen, A.C., Gärtner, K., van Zutphen, L.F.M., 2001. Standardization of animal experimentation (reprinted 2005, 2006). Chapter 5 – In: van Zutphen, L.F.M., Baumans, V., Beynen, A.C. (Eds.), Principles of Laboratory Animal Science. A Contribution to the Humane Use and Care of Animals and to the Quality of Experimental Results, 2nd edition. Elsevier Science Publishers, Amsterdam, The Netherlands, pp. 103–110 (2001).

Bodden, C., von Kortzfleisch, V.T., Karwinkel, F., Kaiser, S., Sachser, N., Richter, S.H., 2019. Heterogenising study samples across testing time improves reproducilibity of behavioural data. Sci. Rep. 9, 8247. https://doi.org/10.1038/s41598-019-44705-2.

Boleij, H., Salomons, A.R., van Sprundel, M., Arndt, S.S., Ohl, F., 2012. Not all mice are equal: welfare implications of behavioural habituation profiles in four 129 mouse substrains. PLoS One 7 (8), e42544. https://doi.org/10.1371/journal.pone.0042544.

Bolivar, V.J., 2009. Intrasession and intersession habituation in mice: from inbred strain variability to linkage analysis. Neurobiol. Learn. Mem. 92 (2), 206–214. https://doi.org/10.1016/j.nlm.2009.02.002.

Budaev, S.V., 2010. Using principal components and factor analysis in animal behaviour research: caveats and guidelines. Ethology 116 (5), 472–480. https://doi.org/10.1111/j.1439-0310.2010.01758.x.

Burokas, A., Arboleya, S., Moloney, R.D., Peterson, V.L., Murphy, K., Clarke, G., Stanton, C., Dinan, T.G., Cryan, J.F., 2017. Targeting the microbiota-gut-brain axis: prebiotics have anxiolytic and antidepressant-like effects and reverse the impact of chronic stress in mice. Biol. Psychiatry 82 (7), 472–487. https://doi.org/10.1016/j.biopsych.2016.12.031.

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

Carreira, M.B., Cossio, R., Britton, G.B., 2017. Individual and sex differences in high and low responder phenotypes. Behav. Process. 136, 20–27. https://doi.org/10.1016/j.beproc.2017.01.006.

Casarrubea, M., Sorbera, F., Crescimanno, G., 2009. Structure of rat behavior in hole-board: I) multivariate analysis of response to anxiety. Phys. Beh. 96 (1), 174–179. https://doi.org/10.1016/j.physbeh.2008.09.025.

Casarrubea, M., Magnusson, M.S., Roy, V., Arabo, A., Sorbera, F., Santangelo, A., Faulisi, F., Crescimanno, G., 2014. Multivariate temporal pattern analysis applied to the study of rat behavior in the elevated plus maze: methodological and conceptual highlights. J. Neurosci. Methods 234, 116–126. https://doi.org/10.1016/j.jneumeth.2014.06.009.

Casarrubea, M., Johnson, G.K., Faulisi, F., Sorbera, F., Di Giovanni, G., Benigno, A., Crescimanno, G., Magnusson, M.S., 2015. T-pattern analysis for the study of temporal structure of animal and human behavior: a comprehensive review. J. Neurosci. Methods 239, 34–46. https://doi.org/10.1016/j.jneumeth.2014.09.024.

Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L., Mogil, J.S., 2002. Influences of laboratory environment on behavior. Nat. Neurosci. 5 (11), 1101–1102. https://doi.org/10.1038/nn1102-1101.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., Home, R., 2005. The use and reporting of cluster analysis in health psychology: a review. Br. J. Health Psychol. 10 (Pt 3), 329–358. https://doi.org/10.1348/135910705X25697. https://doi.org/10.1348/135910705X25697.

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

Crabbe, J.C., Wahlsten, D., Dudek, B.C., 1999. Genetics of mouse behavior: interactions with laboratory environment. Science 284 (5420), 1670–1672. https://doi.org/10.1126/science.284.5420.1670.

Dingemanse, N.J., Dochtermann, N.A., 2013. Quantifying individual variation in behaviour: mixed effect modelling approaches. J. Animal Ecol. 82 (1), 39–54. https://doi.org/10.1111/1365-2656.12013.

Dolnicar, S., Grün, B., Leisch, F., 2016. Increasing sample size compensates for data problems in segmentation studies. J. Bus. Res. 69 (2), 992–999. https://doi.org/10.1016/j.jbusres.2015.09.004.

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

Dutta, S., Sengupta, P., 2016. Men and mice: relating their ages. Life Sci. 152, 244–248. https://doi.org/10.1016/j.lfs.2015.10.025.

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

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

Eisenstein, E.M., Eisenstein, D., 2006. A behavioral homeostasis theory of habituation and sensitization: II. Further developments and predictions. Rev. Neurosci. 17 (5), 533–557. https://doi.org/10.1515/REVNEURO.2006.17.5.533.

Fagerland, M.W., Sandvik, L., Mowinckel, P., 2011. Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables. BMC Med. Res. Methodol. 11 (44), 1–8. https://doi.org/10.1186/1471-2288-11-44.

Ferguson, S.A., Maier, K.L., 2013. A review of seasonal/circannual effects of laboratory rodent behavior. Physiol. Behav. 119, 130–136. https://doi.org/10.1016/j.physbeh.2013.06.007.

Festing, M., 2011. Inbred strains and toxicity testing. In: Kaliste, E. (Ed.), Proceedings of the Eleventh FELASA Symposium and the 40th Scand-LAS Symposium – New Paradigms in Laboratory Animal Science. 14-17 June 2010, Helsinki, Finland. Published by FELASA.

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

Festing, M.F.W., Overend, P., Cortina Borja, M., Berdoy, M., 2016. The Design of Animal Experiments. Reducing the Use of Animals in Research Through Better Experimental Design. Laboratory Animals Handbook No. 14, 2nd edition. Sage Publications Ltd, London, UK.

Finkemeijer, M.A., Langbein, J., Puppe, B., 2018. Personality research in mammalian farm animals : concepts, measures and relationship to welfare. Front. Vet. Sci. 28 (5), 131. https://doi.org/10.3389/fvets.2018.00131.

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

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

Garner, J.P., 2005. Stereotypies and other abnormal repetitive behaviors: potential impact on validity, reliability, and replicability of scientific outcomes. ILAR J. 46 (2), 106–117. https://doi.org/10.1093/ilar.46.2.106.

Gärtner, K., 2012. A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 year long effort to standardize laboratory animals? Int. J. Epidemiol. 41, 335–341. https://doi.org/10.1093/ije/dyr219. Reprint of Lab. Anim. 1990; 24 (1), 71-77, http://doi.org/10.1258/002367790780890347.

Genolini, C., Falissard, B., 2010. Kml: K-means for longitudinal data. B. Comput. Stat. 25 (2), 317–328. https://doi.org/10.1007/s00180-009-0178-4.

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

Guilloux, J., Seney, M., Edgar, N., Sibille, E., 2011. Integrated behavioral z-scoring increases the sensitivity and reliability of behavioral phenotyping in mice: relevance to emotionality and sex. J. Neurosci. Methods 197 (1), 21–31. https://doi.org/10.1016/j.jneumeth.2011.01.019.

Hager, T., Jansen, R.F., Pieneman, A.W., Manivannan, S.N., Golani, I., van der Sluis, S., Smit, A.B., Verhage, M., Stiedl, O., 2014. Display of individuality in avoidance behaviour and risk assessment of inbred mice. Front. Behav. Neurosci. 16 (8), 314. https://doi.org/10.3389/fnbeh.2014.00314.

Jensen, V.S., Porsgaard, T., Lykkesfeldt, J., Henning, H., 2016. Rodent model choice has major impact on variability of standard preclinical readouts associated with diabetes and obesity research. Am. J. Transl. Res. 8 (8), 3574–3584. www.ajtr.org/ISSN:1943-8141/AJTR0023255.

Karp, N.A., 2018. Reproducible preclinical research – is embracing variability the answer? PLoS Biol. 16 (3), e20054113. https://doi.org/10.1371/journal.pbio.2005413. https://doi.org/10.1371/journal.pbio.2005413.

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

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., Altman, D.G., 2010. Improving bioscience research reporting: the ARRIVE guideline for reporting animal research. PLoS Biol. 8 (6), e1000412. https://doi.org/10.1371/journal.pbio.1000412.

Kokras, N., Dalla, C., 2014. Sex differences in animal models of psychiatric disorders. Br. J. Pharmacol. 171 (20), 4595–4619. https://doi.org/10.1111/bph.12710.

Koolhaas, J.M., van Reenen, C.G., 2016. ANIMAL BEHAVIOR AND WELL-BEING SYMPOSIUM: interaction between coping style/personality, stress, and welfare: relevance for domestic farm animals. J. Anim. Sci. 94 (6), 2284–2296. https://doi.org/10.2527/jas.2015-0125.

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

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

Laarakker, M.C., Ohl, F., van Lith, H.A., 2008. 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), 159–184. https://doi.org/10.1007/s10519-007-9188-6.

Laarakker, M.C., van Lith, H.A., Ohl, F., 2011. Behavioral characterization of A/J and C57BL/6J mice using a multidimensional test: association between bloodplasma and brain magnesium-ion concentration with anxiety. Physiol. Behav. 102 (2), 205–219. https://doi.org/10.1016/j.physbeh.2010.10.019.

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

Labots, M., Laarakker, M.C., Ohl, F., van Lith, H.A., 2016. Consomic mouse strain selection based on effect size measurement, statistical significance testing and integrated behavioral z-scoring: focus on anxiety-related behavior and locomotion. BMC Genet. 17 (1), 95. https://doi.org/10.1186/s12863-016-0411-4.

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

Lenth, R., 2019. Emmeans: Estimated Marginal Means, Aka Least-squares Means. R Package Version 1.3.3. https://CRAN.R-project.org/package-emmeans.

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

Magnusson, M.S., 2000. Discovering hidden time patterns in behavior: T-patterns and their detection. Behav. Res. Methods Instrum. Comput. 32 (1), 93–110. https://doi.org/10.3758/BF03200792.

Mogil, J.S., Chanda, M.L., 2005. The case for inclusion of female subjects in basic science studies of pain. Pain 117 (1-2), 1–5. https://doi.org/10.1016/j.pain.2005.06.020.

Nagin, D.S., 1999. Analyzing developmental trajectories: a semiparametric, group-based approach. Psychol. Methods 4 (2), 139–157. https://doi.org/10.1037/1082-989X.4.2.139.

National Research Council, 2003. Guidelines for the Care and Use of Mammals in Neuroscience and Behavioral Research. National Academies Press, Washington.

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

Ohl, F., 2003. Testing for anxiety. Clinic. Neurosc. Res. 3 (4-5), 233–238. https://doi.org/10.1016/s1566-2772(03)00084-7.

Ohl, F., Holsboer, F., Landgraf, R., 2001. The modified hole board as a differential screen for behavior in rodents. Behav. Res. Methods Instrum. Comput. 33 (3), 392–397. https://doi.org/10.3758/BF03195393.

Ohl, F., Arndt, S.S., van der Staay, F.J., 2008. Pathological anxiety in animals. Vet. J. 175 (1), 18–26. https://doi.org/10.1016/j.tvjl.2006.12.013. https://doi.org/10.1016/j.tvjl.2006.12.013.

Paylor, R., 2009. Questioning standardization in science. Nat. Methods 6, 253–254. https://doi.org/10.1038/nmeth0409-253.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., R Core Team, 2018. Nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1.-137. URL:. https://CRAN.R-project.org/package=nlme.

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

Prendergast, B.J., Onishi, K.G., Zucker, I., 2014. Female mice liberated for inclusion in neuroscience and biomedical research. Neurosci. Biobehav. Rev. 40, 1–5. https://doi.org/10.1016/j.neubiorev.2014.01.001.

R Core Team, 2018. R: a Language Environment for Statistical Computing. R foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramos, A., Mormède, P., 1998. Stress and emotionality: a multidimensional and genetic approach. Neurosci. Biobehav. Rev. 22 (1), 33–57. https://doi.org/10.1016/s0149-7634(97)00001-8.

Réale, D., Reader, S.M., Sol, D., McDougall, P.T., Dingemanse, N.J., 2007. Integrating animal temperament within ecology and evolution. Biol. Rev. Camb. Philos. Soc. 82 (2), 291–318. https://doi.org/10.1111/j.1469-185X.2007.00010.x.

Reed, J.M., Harris, D.R., Romero, L.M., 2019. Profile repeatability: a new method for evaluating repeatability of individual hormone response profiles. Gen. Comp. Endocrinol. 270, 1–9. https://doi.org/10.1016/j.ygcen.2018.09.015.

Richter, S.H., 2017. Systematic heterogenization for better reproducibility in animal experimentation. Scand. J. Lab. Anim. Sci. 46 (9), 343–349. https://doi.org/10.1038/laban.1330.

Rodgers, R.J., Dalvi, A., 1997. Anxiety, defence and the elevated plus-maze. Neurosci. Biobehav. Res. 21 (6), 801–810. https://doi.org/10.1016/S0149-7634(96)00058-9.

Rosenthal, R., Rosnow, R.L., 2008. Essentials of Behavioral Research: Methods and Data Analysis. p. 699. Third edition. McGraw Hill, New York. http://nrs.harvard.edu/urn-3:HUL.InstRepos:34945148.

Salomons, A.R., Bronkers, G., Kirchhoff, S., Arndt, S.S., Ohl, F., 2010a. Behavioural habituation to novelty and brain area specific immediate early gene expression in female mice of two inbred strains. Behav. Brain Res. 215 (1), 95–101. https://doi.org/10.1016/j.bbr.2010.06.035.

Salomons, A.R., Kortleve, T., Reinders, N.R., Kirchhoff, S., Arndt, S.S., Ohl, F., 2010b. Susceptibility of a potential animal model for pathological anxiety to chronic mild stress. Behav. Brain Res. 209 (2), 241–248. https://doi.org/10.1016/j.bbr.2010.01.050.

Salomons, A.R., van Luijk, J.A.K.R., Reinders, N.R., Kirchhoff, S., Arndt, S.S., Ohl, F., 2010c. Identifying emotional adaptation: behavioural habituation to novelty and immediate early gene expression in two inbred mouse strains. Genes Brain Behav. 9 (1), 1–10. https://doi.org/10.1111/j.1601-183X.2009.00527.x.

Salomons, A.R., Arndt, S.S., Lavrijsen, M., Kirchhoff Ohl, F., 2013. Expression of CRFR1 and Glu5R mRNA in different brain areas following repeated testing in mice that differ in habituation behavior. Behav. Brain Res. 246, 1–9. https://doi.org/10.1016/j.bbr.2013.02.023.

Sandhu, K.V., Sherwin, E., Schellekens, H., Stanton, C., Dinan, T.G., Cryan, J.F., 2017. Feeding the microbiota-gut-brain axis: diet, microbiome, and neuropsychiatry. Transl. Res. 179, 223–244. https://doi.org/10.1016/j.trsl.2016.10.002.

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

Spruijt, B.M., Gispen, W.H., 1984. Behavioral sequences as an easily quantifiable parameter in experimental studies. Phys. Beh. 32 (5), 707–710. https://doi.org/10.1016/0031-9384(84)90182-3.

Spruijt, B.M., Peters, S.M., de Heer, R.C., Pothuizen, H.H.J., van der Harst, J.E., 2014. Reproducibility and relevance of future behavioral sciences should benefit from a cross fertilization of past recommendations and today's technology: "Back to the future". J. Neurosci. Methods 234https://doi.org/10.1016/j.jneumeth.2014.03.001. 2-1.

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

Tuttle, A.H., Philip, V.M., Chesler, E.J., Mogil, J.S., 2018. Comparing phenotypic variation between inbred and outbred mice. Nat. Methods 15 (12), 994–996. https://doi.org/10.1038/s41592-018-0224-7.

Voelkl, B., Würbel, H., 2019. A reaction norm perspective on reproducibility. bioRxiv, 510941. https://doi.org/10.1101/510941.

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

Wahlsten, D., 2011. Chapter 5: Sample Size in Mouse Behavioral Testing: How to Use Mice in Behavioral Neuroscience, 1st edition. Academic Press, Elsevier Inc., London, U.K. https://doi.org/10.1016/B987-0-12-375674-9.10005-9.

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

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