When numbers become individuals

Improving the quality of behavioral animal experiments by mapping and incorporating inter-individuality in mice

Marloes van der Goot

ISBN
978-94-93270-41-1
DOI
https://doi.org/10.33540/855
Cover design / lay-out
Guus Gijben
Printed by
Proefschrift AIO (proefschrift-aio.nl)

Thesis title inspiration
MSc thesis S. Stolte.

When numbers become individuals

Improving the quality of behavioral animal experiments by mapping and incorporating inter-individuality in mice

Van getal naar individu
Het verbeteren van de kwaliteit van (proef)dierstudies naar (experimenteel) gedrag door het in kaart brengen en meenemen van interindividuele variatie bij de laboratoriummuis

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 10 februari 2022 des middags te 2.15 uur

door

Marloes Hieke van der Goot

geboren op 27 mei 1979
te Arnhem

Promotor:
Prof. dr. S.S. Arndt

Copromotor:
Dr. H.A. Van Lith

With four parameters I can fit an elephant, with five I can make him wiggle its trunk.

John von Neumann

Table of contents

General Introduction

The laboratory mouse is the most prominent experimental vertebrate of choice in bioveterinary and biomedical research. According to a recent estimate, 59% of studies worldwide are conducted with this species in biomedical research alone (Dutta and Sengupta, 2016). The popularity of mice in neurobehavioral preclinical research emerged parallel to increasingly improved quantitative and molecular genetic techniques, such as gene targeting, about 40 years ago (Cryan and Holmes, 2005; Sartori et al., 2011). Over 90% of the human genome is conserved in mice (Guenet, 2005) and being able to fully sequence the murine genome and to introduce genetic modifications allowed researchers to systematically study the neurobiological underpinnings of quantitative traits – such as behavior – and to identify genes and pathways that modulate the expression of these traits at a functional level (Baud and Flint, 2017; Cryan and Holmes, 2005; Crawley, 2000; Hovatta et al., 2005). By now, genetic lines, such as mouse inbred strains and their derivatives, have become the building blocks of quantitative and molecular genetic research (van der Staay, 2006).

Approximately 80% of the published laboratory mouse studies worldwide is conducted with mouse inbred strains (Festing, 2014). Over 450 standard inbred lines of the mouse have been registered, although not all lines qualify as commonly used (Wahlsten, 2011). Mouse inbred strains are produced by successive sister-brother/child-youngest parent mating for at least 20 generations, with the result that an average of at least 99% of the genetic loci are homozygous in every individual mouse (International Committee on Standardized Genetic Nomenclature for Mice & Rat Genome and Nomenclature Committee, 2016). Many strains have been bred however for more than 150 generations, making them essentially homozygous at all loci (Beck et al., 2000). In addition to homozygosity, all animals of the same inbred strain are genetically identical, or isogenic (Festing, 2014). The popularity of mice is also linked to their compact size and cost-effectiveness (i.e. relatively low purchase and maintenance costs per animal) – making this species particularly well suited for the required large sample sizes in for example genetic mapping studies (Wahlsten, 2011).

Standardization, the 3R's, and the third component

The above described developments also made it possible for the field of laboratory animal science to further advance the reduction of variability in quantitative and qualitative biological traits in laboratory animals by means of standardization (Gärtner, 2012). In addition to standardization of environmental conditions, experimental procedures and the state of health of the animals, researchers were now also able to standardize genotype (Gärtner, 2012). Animal experimentation has traditionally been characterized by a strong tendency to reduce any variability between subject animals and environmental factors (Koolhaas et al., 2010). This tradition follows from the generally accepted rationale that uncontrolled variation results in noise that negatively affects the quality and power of animal experiments (Festing, 2004a; Button et al., 2013). The process of standardization may be defined as "the defining of the properties of any given animal (or animal population) and its environment, together with the subsequent task of keeping the properties constant" (Beynen et al., 2003). The objective of standardization is to reduce within-experiment variability (and thereby increase statistical power) and reduce between-experiment variability, which should result in improved comparability of results within and between laboratories (van der Staay and Steckler, 2002; Beynen et al., 2003; Voelkl et al., 2020).

While this phenomenon has been around for many years, the standards that govern which factors should be subjected to standardization are susceptible to change, and have indeed changed over time (van der Staay and Steckler, 2002). For one, the tendency to reduce variability became increasingly tied to the ethical consequences that come with in vivo animal experimentation (Voelkl et al., 2020). After all, reducing within experiment variability increases the power of an experiment to detect a true effect, and therefore reduces the number of

Box I. The 3R's principles: Replacement, reduction and refinement.

The 3R's principles were originally proposed by Russel and Burch (1959) as a framework for humane animal research and are nowadays accepted as the best approach to maximize the quality of experimental results while ensuring the highest standard of ethical consideration (Kirk, 2018). In short, the 3R's principles entail that researchers involved in animal experimentation should at all times strive to avoid or replace the use of animals (replacement), minimize the number of animals used (reduction) and minimize animal suffering and maximize welfare (refinement). The principles function as the regulatory ethical backbone of experimental animal research: they have been embedded in international legislation and regulations in the use of animals in scientific procedures, as well as in the policies of funding bodies (source: www.nc3rs.co.uk/the-3Rs).

animals necessary per study to detect a given effect size (Beynen et al., 2003; Richter et al., 2010; Voelkl et al., 2020). Also, decreasing the variability between experiments reduces the need to repeatedly perform tests and therefore should reduce the number of animals needed (Voelkl et al., 2020). Standardization therefore also became promoted as a means to reduce and refine animal use as required by the 3R's principles (Russel and Burch, 1959; Box I) (Festing, 2004a; 2004b). An unintended side-effect of this conception of standardization however was that in some cases minimizing the number of animals became a goal in itself, in which standardization is applied with the goal to obtain statistically significant results with as little animals as possible (Koolhaas et al., 2010; Voelkl et al. 2020). Also, these mechanisms inadvertently contributed to a conception of standardization that has become synonymous to homogenization of study populations (Koolhaas et al. 2010; Voelkl et al., 2020). This focus on homogenization is interesting, because the above provided definition of standardization by no means implies that all properties of an animal and its environment should be identical (Voelkl et al., 2020). Also, having both too many *and* too little animals negatively affects the quality of experimental results (and with that the required number of animals). A correct implementation of the 3R's principles requires appropriate design and analysis, rather than 'the smallest possible sample size' by definition (Button et al., 2013).

Within this homogenization-oriented concept of standardization one specific source of within and between experiment variability has traditionally been disregarded: inter-individual variability between subject animals, which is here defined as "the collection of behavioral or physiological traits, both innate and acquired, that distinguish one animal from the near relatives that, as far as possible, share the same genetic and environmental background" (Lathe, 2004). Increasing evidence shows however that despite excessive standardization efforts, even experimental animals that share a genetic background, which are kept under standardized husbandry conditions and are subject to standardized experimental protocols may still differ in their quantitative traits to some extent (Koolhaas, et al., 2010; Tuttle et al., 2018; Voelkl et al., 2020). This phenomenon led Gärtner (2012) to suggest that a third component may be at play, (besides genotype and environment) that is not controlled for through environmental and genotypic standardization. In his study, standardization efforts indeed reduced variability related to between group differences (such as genetic background, sex, age) and to a lesser extent by the environment. Random variability (related to within-group variability) however was not affected by standardization efforts (Gärtner, 2012).

The exact constitution of this third component is poorly understood, but a multitude of identified genetic (e.g. Keshavarz et al., 2020; Tam and Cheung, 2020), epigenetic (Lathe, 2004), developmental stochastic (Zocher et al., 2020), environmental (e.g. Crabbe et al., 1999; Freund et al., 2013) and even microbiomic sources (Burokas et al., 2017) indicate that this variability is the result of complex interactions between genetic and environmental factors, that are partly modulated by epigenetic processes (Voelkl et al., 2020; Lathe, 2004). With the complexity of these interactions in mind it is conceivable how increasing standardization efforts have yet failed to control for or remove this source of variability from the equation.

The existence of inter-individual variability between subject animals has always been acknowledged, but for long it was regarded as part of a larger source of unwanted noise, falling within the same category as other sources of extraneous noise (i.e. measurement error, unanticipated environmental effects) that negatively affects the quality and power of animal experiments (Button et al., 2013; Voelkl et al., 2020). The ubiquitous and persistent nature of this phenomenon however, has caused a paradigm shift in recent years, in which the potential of this type of variation is becoming increasingly advocated and appreciated (e.g. Lathe, 2004; Armario and Nadal, 2013; Finkemeijer et al., 2017; Lonsdorf and Merz, 2017; Einat et al., 2018; Karp, 2018; Bello and Renter, 2018; Bushby et al., 2018, Voelkl and Wuerbel, 2019; Voelkl et al., 2020).

Inter-individual variability: Quality and reproducibility of experimental results

For a large part, this appeal emerged in response to persisting issues regarding the quality and reproducibility of experimental results (e.g. Garner, 2005; van der Staay et al., 2010; Richter, 2017; Voelkl et al., 2020; Festing, 2020). The reproducibility of experimental results is one of the cornerstones of good scientific practice, and may be defined as "the ability to produce similar results by independent replicate studies" (Voelkl et al., 2020) or as "the degree of accordance between the results of the same experiment performed independently in the same or different laboratories" (van der Staay et al., 2009). Difficulty with reproducing experimental results between studies has received a surge of attention in recent years, with estimates of inability to reproduce preclinical research findings that range from 75% to 90% (Begley and Ioannidis, 2015). A lack of reproducibility not only undermines scientific advancement and translatability of preclinical results to clinical practice, it again poses severe ethical concerns due to the continued use of laboratory animals being subjected to unjustified pain and distress (Karp, 2018; Voelkl et al., 2020). Not

surprisingly, this reproducibility crisis (Baker, 2016) has also been termed by some scientist the 4th 'R', that should be part of the ethical principles governing animal experimentation (Canadian Council on Animal Care, 2019).

This lack of reproducibility has been associated with a multitude of factors, ranging from issues related to data collection, to data analysis and reporting (Ioannidis, 2005; Karp, 2018). Only recently, variability in phenotypic response entered the picture as another massive factor in the difficulty of translating experimental findings from one lab to another (Voelkl et al., 2020). To better control for this type of variability, and as such improve the quality of animal experimentation, an increasing number of strategies are being proposed that systematically account for phenotypic variation in experimental design and statistical analysis (Richter et al., 2010; Richter, 2017; Kafkafi et al., 2018; Bello and Renter 2018; Voelkl and Wuerbel, 2020; Voelkl et al., 2019). Phenotypic variation in this context should be regarded as the net result of the combined effects of genotype and an individual's response to the environment, integrated over an individuals' lifetime (Voelkl et al., 2020). Within this concept, inter-individual variability is considered part of a larger source of variation that affects differences between individuals and phenotypic plasticity (i.e. the extent to which an individual changes its phenotype in response to the environment, Voelkl et al., 2020). The majority of approaches listed below therefore focus on accounting for phenotypic variation in its broadest sense, and do not exclusively tailor towards accounting for inter-individual differences, but to any potential source of variability, including for example environmental conditions.

One of these approaches is systematic heterogenization, which entails the systematic incorporation of known sources of variability in the design of an animal experiment (van der Staay et al., 2009; Richter et al., 2010; Richter, 2017, Voelkl et al., 2020). According to this view, the risk of a primary focus on reducing within-experiment variation by means of excessive standardization is that this may induce unrealistically low variability, such that obtained results are specific for that study context only – and therefore are difficult to replicate as they do not generalize to the population of interest (i.e. the standardization fallacy; Würbel, 2000; Voelkl et al., 2021). By systematically incorporating known sources of phenotypic variation in a single-lab study, this study is assumed to better represent the range of variation that one may find between experiments, which in turn should improve the generalizability, and reproducibility of results (van der Staay, 2006; Richter et al., 2010; Voelkl et al., 2020). Sources of variability that may be systematically heterogenized within a study are for

example genotype, sex and age, but also housing or test condition (i.e. testing time, experimenter) (van der Staay et al., 2009; Bodden et al., 2019; Richter, 2020; Voelkl et al., 2020).

Second, using the appropriate experimental design may also reduce phenotypic variation and improve the quality of experimental outcomes (Festing, 2020). A common design in preclinical animal experimentation, is the 'randomization to treatment group' design, in which subjects (experimental units) are randomly assigned to treatment groups, but the order in which the experiment is done is not randomized (Festing, 2020). A potential risk of this design is that treatment groups differ in micro-environment, time of test etc., which may lead to environmental effects that may be mistaken for treatment effects (Festing, 2020). According to Festing (2020) only two designs are appropriate for widespread use in preclinical animal research: Completely randomized factorial designs (CR), and complete randomized block designs (CRB).

In the first, each subject is randomly assigned to a treatment group, and subjects receiving different treatments are randomly intermixed in the experimental environment (Festing, 2020). Factorial designs are of great value as these designs allow one to increase the generality of the findings by simultaneous testing of two or more factors (e.g. sex, strain) without having to increase the overall sample size (Festing, 2016). Such designs may consist of between-group factors, within-group factors or a mixed factorial design, which combines between- and within-group factors. In a CRB, the experiment is split up into a number of independent blocks or 'mini experiments', with a single individual (randomly) assigned to each of the treatments (Festing, 2016; Festing, 2020). Within each block experimental conditions may be rigorously standardized (for example subjects are as similar as possible in age, weight, gender etc.), so that differences in response are likely attributable to the treatment. Between blocks however, conditions may vary, and block is included as a random effects factor in the analysis. This design provides additional control over environmental conditions affecting phenotypic variation, and is therefore more powerful than a CR design (Festing, 2020).

Third, one may also account for phenotypic variation in statistical models, for example by including the interaction between laboratory and genotype as a random factor in multilevel models (Kafkafi et al., 2018). Another effective strategy may be to determine what amount of variation in the data is related to phenotypic variation. In behavioral ecology, and to a lesser extent in preclinical

research, increased interest in individual differences within the context of animal personality research has resulted in the development of statistical frameworks that facilitate the analysis of such variance (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 which different variance components related to individual variation can be quantified (e.g. Dingemanse and Dochtermann, 2013; Alleque et al., 2017; Bushby et al., 2018). Examples of these components are inter-individual variability (differences in response between individuals of the same study population) or intra-individual variability (variability in response measured within the same individual).

Inter-individual variability: Translational value of animal models of neurobehavioral disorders

In concurrence with the increased interest in inter-individual variability in the context of the reproducibility of experimental results, this type of variation became increasingly recognized as an important factor in preclinical animal models of neurobehavioral disorders and psychopathologies (Armario and Nadal, 2013; Einat et al., 2018). In humans, the susceptibility to develop psychopathologies and the response to treatment is known to vary greatly between patients (Einat et al., 2018). In a clinical setting, exposure to similar conditions may result in the development of psychopathologies in some, while others remain unaffected (Kazavchinsky et al., 2019). Einat et al. (2018) therefore argued that incorporation of this variability in animal models may make these models more representative and homologous. In addition, assessment of such variability in response to stressors, and/or treatment could improve our understanding of the underlying biological mechanisms that affect differential susceptibility or sensitivity to treatment in humans (Einat et al., 2018).

It has been argued that a first step to assess differential susceptibility to develop a particular disorder is to provide an in-depth, individual based characterization of baseline behavior (Armario and Nadal, 2013). One could then subsequently assess whether differential baseline levels affect variability in response to treatment (Armario and Nadal, 2013). Such in-depth characterization may then be followed up by the identification of biological markers underlying differential expression of a particular behavior, with the ultimate goal of

establishing neurobiological correlates of that particular disease (Armario and Nadal, 2013). Putative genes driving differential susceptibility may be studied in parallel (Armario and Nadal, 2013).

In this field, several strategies exist to harness inter-individual variability in response to treatment or test conditions (Harro, 2010). A common approach for example is to exploit already available strain differences (Armario and Nadal, 2013). Some strains may consistently rank low on the expression of a particular phenotype (for example anxiety-related behavior) while others are predominantly found at the other end of the spectrum (i.e. Trullas and Skolnick, 1993; Bothe et al., 2005; Moy et al., 2007; Tam and Cheung, 2020). Selective breeding strategies form a more active, targeted approach to addressing inter-individual variability. By means of selective breeding, genetically divergent lines of animals that differ on a particular phenotype are produced (Harro, 2010). Well-known examples include the genetically selected lines of high anxiety and low anxiety behavior (HAB, LAB) from outbred WIST rats (Wistar) and CD-1 mice (Liebsch et al., 1998; Kromer et al., 2005; Landgraf et al., 2007; Harro, 2010). Another genetic strategy involves genetic manipulation of targeted genes, by (temporarily) silencing or over-expressing a particular gene (Armario and Nadal, 2013). The advantage of these genetic strategies is that they provide a reliable genetic basis and ensure a stable expression of the trait of interest. A disadvantage of these approaches however is that strains may become subject to genetic drift (Taft et al., 2006; Harro, 2010).

These strategies have fundamentally improved our understanding of differential expression in behavioral dysfunction and the underlying mechanisms that modulate them, and as such are indispensible (Einat et al., 2018). In the majority of these studies however, the interpretation of experimental results is still centered around the comparison of group effects, in which numerical data are summarized by their measures of central tendency and associated measures of dispersion (Einat et al., 2018). When describing the data this way, individual values are not taken into consideration. Also, these strategies do not tailor towards accounting for any potentially confounding influence of inter-individual variability. A frequently employed approach that does zoom in on inter-individual variability includes the use of simple selection strategies to distinguish between differential subpopulations within an experimental pool (Harro, 2010). This approach focuses on separating animals whose expression of a particular trait lies on the outer ends of the distribution, for example by means of a median/tertile/quartile split (Harro, 2010; Einat et al., 2018). Individuals

that lie on the outer end of a distribution may be labeled as either high or low responders, while the remaining individuals may either be included as controls or disregarded for further analysis (see Harro, 2010 for examples related to inter-individual differences in animal models of affect). The advantage of this approach over genetic strategies is that it is more flexible and allows for incorporation of both genetic and epigenetic mechanisms (Harro, 2010).

Interestingly, it has been demonstrated in some of these studies, that the existence of subpopulations may hamper the detection of significant differences at group level (i.e. increase the chance on a Type II error, Barbelivien et al., 2008; Lonsdorf and Merz, 2017). These findings indirectly suggest that considering the occurrence of inter-individual variability in an experimental pool may benefit the quality of experimental results. A disadvantage of this approach however is that in most studies, inter-individual variability is established by means of an artificially predetermined quantile. Such criteria may lead to a loss of resolution and power, for example in a median split, where every value below and above the median is considered equal, regardless of its position on the phenotypic distribution (Irwin and McClelland, 2003). Also, in the case of a tertiary or quartile split only the outer ends of a study population are compared (Stegman et al., 2019), instead of the entire phenotypic spectrum.

These criteria contrast with the generally accepted conception of human psychopathology being a continuum from health to pathology (Insel et al., 2010). Such a conceptualization warrants the exploration of subgroups across the entire study population, and not just the outer ends. Ideally thus, identification of subgroups should be based on any variability in the data, rather than a predefined criterion (Lonsdorf and Merz, 2017). In the present thesis, we therefore assessed inter-individual variability between subject animals using a data-driven, continuum-based analysis approach that allowed us to explore the existence of meaningful subpopulations on the basis the variability in the data itself, across the entire phenotypic spectrum. We assessed this variability in multiple mouse inbred strains, within the context of one of the most prevalent human neurobehavioral disorders: Anxiety disorders.

By using this approach, we defined experimental animals on an individual level with a two-fold purpose: i) to map inter-individual variability in behavioral and physiological responses within multiple inbred strains of mice and ii) to explore whether systematic incorporation of this inter-individual variability in the design of an animal experiment would affect the quality of experimental results.

Anxiety

Anxiety disorders in humans

Anxiety is a basic emotion that is highly conserved through evolution (Ohl et al., 2008). In humans, anxiety is generally defined as a (sustained) response to distress in anticipation of a (potential) future threat (DSM-V 2013, Taylor and Wahlen, 2015). It differs in that respect from the closely related concept of fear, which is typically defined as a transient emotional response to a specific stimulus or event (DSM-V 2013, Taylor and Wahlen, 2015). In the neurosciences, this distinction is defined as the response to an undetermined, potentially dangerous situation (anxiety) versus the response to an explicit hazard (fear) (Sartori et al., 2011). Anxiety in humans is often perceived as an emotionally negative, unpleasant experience (Beck et al., 1985). From an evolutionary perspective however, this construct is a biologically useful and adaptive response, warning individuals about potential danger and initiating appropriate somatic, cognitive, emotional and behavioral responses that avoid harm and thus promote survival (Sartori et al., 2011).

Anxiety responses may become pathological when they become persistent, uncontrollable, excessive and inappropriate and are even elicited in absence of a recognizable threat (Ohl et al., 2008; Sartori et al., 2011). If such responses persist for a period longer than 6 months these are classified as pathological (DSM-V, 2013). Anxiety disorders are a class of psychiatric disorders that are characterized by elements of excessive fear and anxiety, and behavioral disturbances (DSM-V, 2013) and are the most prevalent psychiatric disorders that affect nearly 30% of adults at some point in their lives (Bandelow and Michaelis, 2015). This high prevalence, along with the associated high health care and economic costs as well as the lower health-related quality of life, emphasizes the importance of increasing our understanding of how these disorders manifest and may be treated (Bandelow and Michaelis, 2015; Taylor and Whalen, 2015). The etiology of anxiety disorders is part associated with genetic factors, with an estimated 30-40% of heritability (Hettema et al., 2001), but also environmental factors such as negative experiences and early life stress (Merikangas and Pine 2002; Young et al., 2008). These factors interact, resulting in differential susceptibility between individuals to develop an anxiety disorder (Wermter et al., 2010).

Measuring pathological anxiety in mice

Mouse models of anxiety enable the study of brain-behavior relations, with the purpose of increasing understanding of normal and abnormal human behavior and the neuronal and neuro-endocrinological processes that modulate it: "An animal model with biological and/or clinical relevance in the behavioral neurosciences is a living organism used to study brain-behavior relations under controlled conditions, with the final goal to gain insight to, and to enable predictions about, these relations in human and/or a species other than the one studied, or in the same species under conditions different from those on which the study was performed" (van der Staay, 2006).

Measuring anxiety-related behavior in mice is mostly conducted in test procedures in which subjects are exposed to unfamiliar aversive places or stimuli (Belzung and Griebel, 2001). Behavioral tests that measure anxiety in mice can be roughly divided in paradigms measuring conditioned responses to aversive and often painful events (e.g. exposure to electric foot shock or predator scent) and more ethologically based paradigms that measure unconditioned, innate/natural expression of anxiety-related behavior (Ohl, 2003). Examples of the latter are the Open Field (OF), Light-Dark Box (LD), the Elevated Plus Maze (EPM), the Elevated Zero Maze (EZM, Belzung and Griebel, 2001) and the modified Hole Board (mHB) (Ohl, 2001a; Labots et al., 2015). Figure 1 presents a visual and schematic representation of the first four assays. The modified Hole Board is discussed in more detail in a separate section of this chapter. The advantage of unconditioned behavioral tests is that they are less susceptible to potentially confounding interference of learning, hunger, or nociceptive mechanisms and allow for a truly comprehensive behavioral profiling (Rodgers et al., 1997). A detailed overview and description of different conditioned and unconditioned tests of anxiety can be found in Ohl (2003) and Belzung and Griebel (2001).

Research has shown that mice of the same inbred strain may differ in their expression of unconditioned anxiety (Ducottet and Belzung, 2004; Cohen et al., 2008; Jakovcevski et al., 2008; Lewejohann et al., 2011; Freund et al., 2013; Keshavarz et al., 2020). Cohen et al. (2008) for example showed that individuals within the same inbred strain differed in unconditioned, baseline anxiety-levels (measured in the elevated plus maze), with individuals within DBA, C57BL/6, NZB and SJL being classified as low and moderate anxious. Jakovcevski et al. (2008) showed that C57BL/6 mice could be divided in two subpopulations exhibiting high and low trait anxiety, when measuring the latency to enter

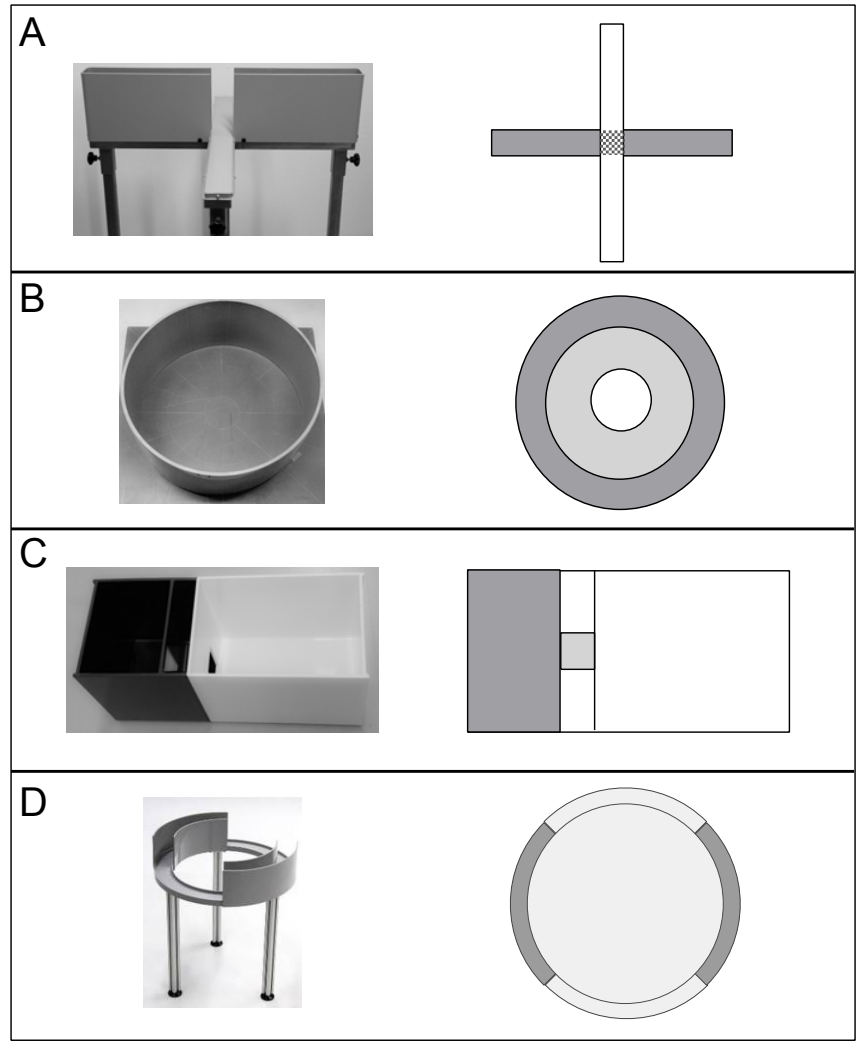


Figure 1. Commonly used behavioral tests of unconditioned anxiety-related behavior, represented by a picture (left) and a structure diagram (right). Grey areas in each structure diagram represent 'protected' zones and white areas represent 'unprotected' zones. (A) The Elevated Plus Maze (EPM): an elevated cross shaped runway with two enclosed arms (grey in the schematic overview) and two open arms (white). The arms are connected by a central platform (grid lines). (B) the Open Field (OF): a circular assay that distinguishes between the center (white), the middle (light grey) and the periphery (dark grey). (C) the Light-Dark Box (LD): Consists of a light and dark compartment, separated by a tunnel. (D) The Elevated Zero Maze (EZM): an elevated circular runway with alternating enclosed areas (grey) and open light areas. Pictures in A, B, C courtesy of Anne-Marie Baars. Picture in D obtained from www.tse-systems.com/product-details/elevated-plus-zero-maze. Figure modified after Labots (2017); Chpt 1, pp. 23).

an unfamiliar arena from their home cage. Acute stress enhanced anxiety responses in a follow up test in individuals with high anxiety, but not in individuals classified as having low anxiety. Furthermore, Ducottet and Belzung (2004) found two distinct profiles of high and low emotional reactivity within a population of BALB/cJ and C57BL/6J strains. Mice characterized as highly emotional reactive combined high avoidance towards new or aversive areas of an environment with low levels of exploratory behavior, while mice with a low reactivity profile displayed high levels of exploration towards novel/ aversive areas. Lastly, inter-individual variability has also been demonstrated in differential exploratory and locomotor levels, with C57BL/6 individuals differing in their activity patterns when recorded in the home environment (Freund et al., 2013) or in an open field test (Vidal-Gomez, 2016).

In humans as well as rodents however, anxiety comprises both innate (trait) and situation-evoked (state) anxiety (Ohl, 2003). State anxiety reflects the direct response to an anxiogenic stimulus whereas trait anxiety is considered an enduring feature of an individual (Belzung and Griebel, 2001). In mice, these two forms have been found differentially sensitive to pharmacological compounds (Belzung and Griebel, 2001). Trait anxiety has been proposed as a potential indicator of pathological anxiety in rodents (Belzung and Griebel, 2001). The aforementioned tests of anxiety-related behavior in mice almost always evaluate situation-evoked behavior (i.e. state anxiety), reflecting a 'normal, adaptive' anxiety response, which makes it difficult to measure, and model, trait anxiety in animals (Ohl et al., 2008). A common strategy therefore to study trait anxiety is to assess behavior and physiology in mouse models that are characterized by long-term enhanced anxious states, through genetic selection (for example inbred strains with inborn enhanced anxiety) or through manipulation of environmental conditions (i.e. adverse rearing, chronic stress exposure) (Belzung and Griebel, 2001; Sartori et al., 2011). However, individuals of such genetic lines with high trait anxiety often show a tendency to also show high state anxiety and vice versa, making it difficult to make a clear delineation between the two (Ohl, 2003).

Ohl and colleagues therefore proposed an additional approach to assess pathological anxiety in animals, one in which the adaptive nature of anxiety responses is emphasized, rather than the enduring feature of trait anxiety. They defined pathological anxiety as "a persistent, uncontrollable, excessive, inappropriate and generalized dysfunctional and aversive emotion, triggering physiological and behavioral responses lacking adaptive value. Pathological

anxiety-related behavior is a response to the exaggerated anticipation or perception of threats, which is incommensurate with the actual situation" (Ohl et al., 2008). This concept and definition relies heavily the rationale that 'state' anxiety responses (i.e. anxiety responses that are measured in conventional behavioral anxiety tests) may be considered adaptive as they enable individuals to respond appropriately to such stimuli. Such an adaptive, 'normal anxiety' response should then be characterized by a decrease in anxiety-related behavior during repeated or prolonged exposure to the stressor. In a series of studies, Ohl and colleagues investigated whether the opposite of such a response, i.e. enhanced anxiety during repeated exposure, may then in turn mirror a non-adaptive, pathological anxiety response (Ohl et al., 2008; Boleij et al., 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013).

In the studies by Ohl and colleagues, the adaptive quality of anxiety responses was assessed by means of the behavioral habituation and sensitization of anxiety-related behavior. These two contrasting forms of non-associative learning are defined as the decremental (habituation) and 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). These studies assessed behavioral habituation of anxiety-related behavior in BALB/c and a number of substrains of 129 mice and consistently found that these strains displayed contrasting behavioral profiles when repeatedly exposed to an initially novel environment (Boleij et al., 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). BALB/c mice, an inbred strain known for its highly anxious phenotype (Bothe et al., 2005), displayed high initial levels of anxiety-related behavior that decreased as trials progressed (Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). At the same time low levels of exploration and locomotion increased, indicating a successful adaptation to the test. In contrast, several sub-strains of 129 mice displayed low levels of anxiety-related behavior during initial exposure that increased as trials progressed, while activity levels remained low or decreased, reflecting a failure to habituate their anxiety response (Boleij et al., 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). These contrasting behavioral profiles were modulated by administration of anxiolytic compounds (BALB/cJ: Diazepam; 129P3/J: MPEP; Salomons et al., 2012). In addition, they were associated with strain differences in neuronal activity (measured as c-Fos

expression) in the prelimbic cortex and dentate gyrus (Salomons et al., 2010a; 2010c), regions that are involved in higher cognitive processes that regulate the stress response (McEwen, 1999). C-Fos expression was increased in BALB/cJ after exposure to the mHB, but not in 129P3/J mice. This effect was reversed in 129P3/J after administration of 2-methyl-6-(phenylethynyl)pyridine (MPEP; Salomons et al., 2012). This impaired neural processing between the prelimbic cortex and emotional brain areas (amygdala) was regulated by impairment of the corticotropin releasing factor CRF1 (Salomons et al., 2013). It was concluded that repeated exposure to the same test set up allows for identification of adaptive and non-adaptive phenotypes when comparing different inbred strains of mice, and that 129 mice display a non-adaptive phenotype that may be a potential indicator of pathological anxiety.

These differential profiles however were based on comparison of (sub-) strain means and medians. Retrospect analyses on these studies indicated that the variation in anxiety-like responses was quite substantial, as it was not uncommon to find coefficients of variation over 100%. Perhaps the third component played a role here as well. In this thesis we expanded on this series of studies to map and characterize any inter-individual variability in adaptive capacities of BALB/c and 129 mice. In addition, we extended this assessment to another commonly used mouse inbred strain: C57BL/6.

Behavioral and physiological indicators of anxiety in mice: A multidimensional construct

In natural animal populations, accumulating evidence suggests that phenotypic variation is not infinite but rather clustered around a limited number of phenotypes (Koolhaas et al., 2010; Koolhaas and van Reenen, 2016). In rodents for example, often specific traits are correlated with other traits, for example in coping style (Boissy 1995; Koolhaas and van Reenen, 2016). Coping style may be defined as "a coherent set of behavioral and physiological responses to challenge that is consistent over time and across contexts that is characteristic for a certain group of individuals" (Koolhaas et al., 1999). The integration of multiple measures as a means to assess individuality is therefore becoming increasingly advocated in the case of complex behavioral and/or physiological constructs, such as coping style, emotionality, temperament etcetera (Hager et al., 2014; Ramos and Mormède, 1998). Koolhaas and van Reenen (2016) for example proposed the combined assessment of coping style, emotionality and sociality to assess individual vulnerability to stress related diseases. Likewise, Réale et al. (2007) proposed the combined analysis of traits to describe the full

nature of temperament. Anxiety-related behavior in mice is an equally complex construct that involves the expression of both anxiety-related and activity behaviors (Ohl, 2003, Griebel et al., 2000; Bothe et al., 2005). In this thesis, we therefore used a multidimensional approach to assess inter-individual variability in habituation of anxiety responses. The remainder of this section describes the different behaviors that comprise the expression of anxiety in mice in more detail.

A central feature of anxiety-related behavior in mice is the tendency to avoid potentially harmful stimuli such as an unprotected area, with high avoidance behavior reflecting high anxiety (Barnett, 1975; Belzung and Le Pape, 1994). When first exposed to a novel environment, animals typically start exploring the environment to reduce uncertainty and to gather information about resource availability (e.g. Tebbich et al., 2009). Rodents for example, explore a novel environment along the edges, while avoiding the central, i.e. unprotected area (Ohl, 2003). The aversive nature of such an area may be enhanced by increased illumination levels, which increases predation risk (e.g. Vásquez, 1994) or by elevating it and enabling the animal to see the edge (Ohl, 2003). Avoidance behavior has been shown sensitive to compounds that exert an anxiolytic effect in humans, although this may differ between types of behavioral tests (i.e. Belzung and Berton, 1997; Bourin, 2007). Although it is one of the most prominent indicators of anxiety, the expression of this behavior may be modulated by locomotion and exploration behavior, which should be taken into account for a correct interpretation (Ohl, 2005). More on this interplay is described below.

Risk assessment behavior constitutes a second species-specific expression of anxiety-related behavior (Rodgers et al., 1997; Blanchard et al., 1993). In mice this behavior is characterized by stretched attends and directed sniffing, and is aimed at gathering information by carefully approaching a potentially threatening stimulus and/or by scanning the surrounding area (Ohl et al., 2008). Risk assessment is believed to be an active defense pattern, making it closely related to anxiety (Blanchard et al., 1993). Although closely related, risk assessment has been found to represent a dimension independent of avoidance behavior (Ohl et al., 2001; Rodgers and Johnson, 1995; Laarakker et al., 2008; Labots et al., 2016). It has also been suggested to be a more sensitive indicator of anxiety-related behavior than avoidance behavior (Ohl et al., 2008). Display of risk assessment is however highly strain specific (O'Leary et al., 2013), making this behavior difficult to compare across inbred strains. Arousal-related

behaviors have also been associated to anxiety in mice, particularly grooming (Estanislau, 2012; Kahlueff and Tuoimaa, 2005) and defecation/urination (Turri et al., 2001). Like risk assessment, these indicators of anxiety are highly strain specific (O'Leary et al., 2013).

As briefly pointed out above, the expression of anxiety-related behavior is strongly interlinked with activity behavior: locomotor and exploration activity. Exploration comprises behaviors such as rearing, sniffing and climbing (Ohl, 2005). Most tests that measure unconditioned anxiety capitalize on the motivational conflict between the motivation to explore a novel environment and the drive to avoid potentially harmful stimuli that arises when rodents are introduced to a novel environment (the approach/avoidance conflict, Ohl, 2003; Armario and Nadal, 2013; O'Leary et al., 2013). When exposed to such an environment, exploration may thus be gradually inhibited by anxiety. Exploration is therefore regarded as an indirect measure of anxiety, and is often negatively correlated with avoidance behavior. Administration of anxiolytic compounds has been shown to reverse this inhibition of exploratory behavior (Rodgers et al., 1992; Belzung and Berton, 1997). Similarly, differences in locomotor activity may confound the interpretation of avoidance behavior as an indicator of anxiety, as a lack of exploration of an unprotected area may just as well be the result of overall low levels of activity (Ohl, 2003). Locomotor activity comprises horizontal locomotion and levels of locomotion may differ strongly between strains, with some strains being qualified as highly active, while other strains are known for their low activity levels (O'Leary et al., 2013). Qualitative analyses have demonstrated that avoidance behavior is not merely the lack of locomotor activity when mice are exposed to a novel environment (Lister, 1987; Rodgers and Johnson, 1995; Laarakker et al., 2008; Labots et al., 2016). A comprehensive interpretation of anxiety-related behavior in such a setting should however ideally always include an account of locomotor activity, either by concurrent assessment of this behavior (**Chapter 2, 3, 4, 5**) and/or by including this behavior in the analysis as a covariate (**Chapter 5**).

Finally, the incorporation of physiological indicators of anxiety is important when studying anxiety-related behavior as the expression of this behavior is in part influenced by physiological processes (Cockrem, 2013). The secretion of glucocorticoids in the blood via the hypothalamo-pituitary-adrenal (HPA) axis mediates the response to physiological and emotional stressors, allowing rodents (and other species, including humans) to respond appropriately to a situation (Cockrem, 2013; Ebner and Singewald, 2017). In the context of anxiety,

this response facilitates increased attention and risk assessment towards potentially threatening stimuli (McNaughton and Corr, 2004). In rodents, the primary glucocorticoid hormone, corticosterone, is often used as a measure of acute stress (Herman et al., 2016). Circulating corticosterone levels have indeed been repeatedly associated with anxiety-related behavior (Korte et al., 2001; Ardayfio and Kim, 2006), such as risk assessment (Rodgers et al., 1999). Similar to inter-individual differences in the behavioral expression of state anxiety described above, circulating corticosterone levels have also been demonstrated to differ between individuals (Sgoifo et al., 1996; Rougé-Pont et al., 1998; Cockrem, 2013; Ebner and Singewald, 2017; Weger and Sandi, 2018).

Multidimensional nature of anxiety: genetic factors

The complex, multidimensional nature of anxiety-related behavior is also reflected on a genetic level. Research focusing on the genetic underpinnings of anxiety-related behavior in both humans and mice has demonstrated that anxiety-related behavior has a complex inheritance, with multiple genes interacting with each other as well as with epigenetic and environmental factors (Clément et al., 2002). These complex interactions thereby modulate inter-individual differences in susceptibility to develop an anxiety disorder (Wermter et al., 2010). The identification of candidate genes and pathways that contribute to these differences therefore is one of the key goals in neuro-behavioral preclinical research (Baud and Flint, 2017). The identification of candidate genes of anxiety-related behavior typically starts with the mapping of quantitative trait loci (QTLs): the most likely region(s) of a chromosome that is/are associated with genetic variation for a particular trait (Lander and Botstein, 1989; Labots et al., 2016; Baud and Flint, 2017). Chromosome substitution strains (CSSs) constitute one of the available genomic resources that can be used to detect such QTLs (Nadeau et al., 2012). CSSs are particularly suitable for complex trait analysis as the complex gene is partitioned in a defined and reproducible manner (Nadeau et al., 2012). CSSs are created by transferring a single chromosome from one inbred strain (the donor strain) onto the genetic background of another inbred strain (the host strain) by repeated backcrossing (Singer et al., 2004; Nadeau et al., 2012), see Figure 2.

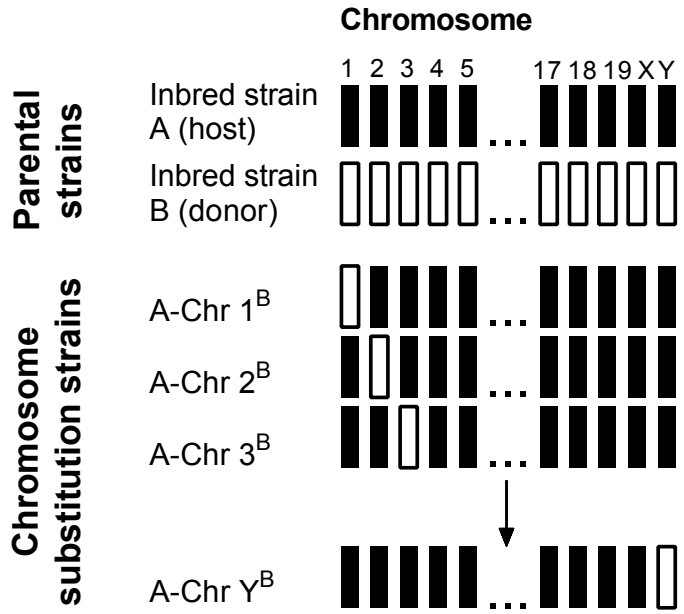


Figure 2. Schematic overview of a chromosome substitution strain. The full length chromosome from a donor inbred strain B is transferred onto the genetic background of a host inbred strain A (Image modified after Laarakker (2009); Chpt 1, pp. 20).

The first complete mouse CSSs panel was created from C57BL/6J and A/J strains (Nadeau et al., 2012). Which are frequently compared in anxiety research because of their contrasting anxiety phenotypes (Trullas and Skolnick, 1993; Bouwknecht and Paylor, 2002; Laarakker et al., 2011). C57BL/6 are known for their low anxious, highly active behavioral profile while A typically display high levels of anxiety related behavior and low levels of activity (e.g. Bolivar et al., 2000; Bothe et al., 2005; Laarakker et al., 2011; Tam and Cheung, 2020). QTL-analyses using C57BL/6 and A in different mapping populations have identified several chromosomal regions that were associated with the expression of unconditioned anxiety-related behavior, with a prominent role for QTLs on chromosomes 1, 10, 15 and 19 (see Laarakker et al., 2008, Table 11 for a comprehensive overview of this literature). The majority of these studies however measured only the magnitude of an unconditioned anxiety response (i.e. the response to a one time exposure to a stressor). In the present thesis, we explored to what extent the chromosomal regions associated with this one-off response also play a role in the modulation of the change of anxiety-related behavior over time (**Chapter 5**).

Modified Hole Board

A behavioral test for unconditioned anxiety that is especially suitable for the phenotyping of complex constructs is the modified Hole Board (mHB). The mHB was developed from the notion that only a rich test environment will allow animals to express their rich behavioral repertoire (Ohl, 2005) and combines characteristics of an open field, a hole board and a light dark box (Ohl, 2001a). As such it is allows for the assessment of multiple behavioral constructs, in contrast to other widely employed test that focus on a single, or a limited number of behavioral parameters (Ohl 2001a; Labots et al., 2015). Other advantages include that this design allows for a reduction in the number of animals as well as the time needed for testing and thereby overcomes the disadvantages of a test battery. In addition to this reduction, the mHB circumvents possible test order effects and the risk that experiences from one test carry over to the next assessment. A drawback of the currently available mHB is that it requires manual scoring of behavior, which is labor intensive and which requires more effort to reduce potential handling and experimenter/observer effects on the outcomes compared to, for example, an automated scoring system. In addition, one should control for order effects if the group compartment of the mHB is used, but this was not the case in the present thesis.

The mHB measures a wide range of behaviors, such as avoidance behavior, risk assessment, arousal, exploration, locomotor activity, social affinity and cognition (Labots et al., 2015). It has been pharmacologically validated for mice and rats (Ohl et al., 2001a; 2001b), and is primarily used for phenotyping in these two species. The mHB consists of a grey PVC opaque box (100 x 50 x 50 cm) with a board made of the same material (60 x 20 x 20 cm) functioning as an unprotected area, as it is positioned in the center of box. This test compartment is separated from an additional compartment (50 x 50 x 50 cm) in which group mates of the experimental animal may be placed. This group compartment is separated from the test compartment by a transparent, perforated PVC divider (Ohl et al., 2001a). If the group compartment is not used, an opaque PVC divider is placed to separate the two compartments (Chapters 2-5). As depicted in Figure 3, the board stacks 20 cylinders (diameter 15 mm) in three lines. The area around the board is divided into 10 rectangles (20 x 15 cm) and 2 squares (20 x 20 cm). The aversive nature of the central area may be modulated through increased/decreased light intensity on this part of the box. In addition, a novel and familiar object (for example a bolt and a die) may be placed in the test compartment prior to start of the test to assess object recognition. The familiar object in this situation is then placed in the animals' home cage 2 days before the experiment (Labots et al., 2015).

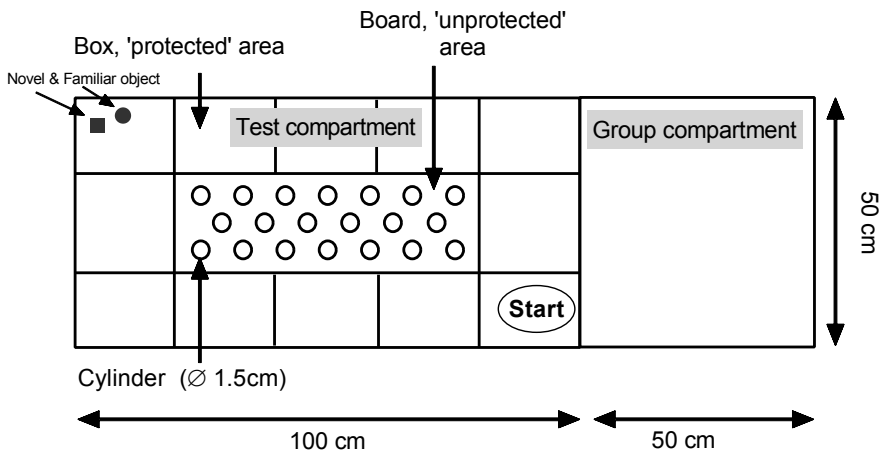


Figure 3. Schematic overview of the mHB.

The mHB thus allows for the assessment of both anxiety-related and activity behavior. Previous research showed that behavioral variables measured in the mHB may be summarized to five behavioral dimensions: avoidance behavior, risk assessment, arousal, exploration and locomotion (Labots et al., 2018; Labots et al., 2016). The dimensions avoidance behavior, risk assessment and arousal may subsequently be combined to a single score representing the motivational system 'anxiety' (Labots et al., 2018; Labots et al., 2016). Especially in unconditioned tests of anxiety-related behavior, it is common to collect a multitude of behavioral variables that reflect different aspects of the same behavior (Ennaceur and Chazot, 2016). In the mHB for instance 35 variables related to anxiety-related and activity behavior may be measured (Laarakker et al., 2008). These variables reflect different aspects of the same behavior, for example frequency, duration and latency of a behavior. Through composite variables these separate scores can be summarized into more manageable and meaningful information, making them more suitable for a multidimensional assessment of anxiety-related behavior as is the objective in this thesis. Another advantage of such composite variables is that they may control for false positives, or a Type I error (Song et al., 2013).

There are different methods to obtain composite variables, and one such approach is integrated behavioral z-scoring. This procedure was originally proposed by Guilloux et al. (2011) and extended by Labots et al. (2018) as a multidimensional approach for behavioral phenotyping in mice. Details

of the rationale, its advantages over other multivariate approaches and the exact procedure may be consulted in Labots et al. (2018). In short, it entails that separate variables that measure different aspects of the same behavior are z-transformed, which measures the amount of standard deviations an observation is above or below the mean of a reference group (Labots et al., 2018). These z-transformed variables are subsequently combined to a single score representing that particular behavioral dimension or motivational system. In the present thesis, behavioral scores obtained in the mHB were summarized to the aforementioned five behavioral dimensions (**Chapters 2, 3, 4 and 5**) and the motivational system 'anxiety' (**Chapter 5**). We used the pooled data as a reference group in all chapters, as suggested by Labots et al. (2018).

Statistical methods

As outlined at previously in this chapter, our objective was to use a multidimensional, data-driven approach to assess inter-individual variability in habituation of anxiety-related behavior in mice. By doing so we aimed to explore whether we could identify subpopulations of mice that would share a similar temporal response across multiple behavioral dimensions: so-called response types. This put two requirements on the anticipated statistical approach that would be suitable for this purpose: (i) it should facilitate individual-based assessment of behavioral and/or physiological response trajectories (i.e. longitudinal, repeated measures data) and (ii) it should do so over multiple behavioral dimensions at the same time (i.e. facilitate multivariate assessment).

With respect to the first, a defining feature of the previously described statistical frameworks that facilitate the analysis of individual variation (section "*Inter-individual variability: Quality and replicability of experimental results*") is that they summarize the different variance components related to individual variation to single point data. This also holds for individual variation in longitudinal responses and as such they unfortunately do not allow for the assessment of the shape or progression of individual response curves. Individual-based trajectory analysis is facilitated by other model-based approaches however, such as mixture modelling techniques and latent class growth analysis (LCGA) (Genolini and Falissard, 2010). LCGA for example was used by Galatzer-Levy et al. (2013) to identify differential temporal profiles of threat extinction learning within a population of outbred rats. The advantage of LCGA and mixture modelling is that formal tests can be used to evaluate the validity of the partitioning of the data (Genolini and Falissard, 2010). Unfortunately

however they do not tailor towards our second requirement: the concurrent assessment of individuality on multiple behavioral dimensions, or a multivariate approach. Multidimensionality of behavioral constructs in turn is often assessed with multivariate techniques such as principal component analysis (PCA) and exploratory factor analysis (i.e. Feyissa et al., 2017; Laarakker et al., 2008; Labots et al., 2016). These classification methods however only facilitate the analysis of single point data, and not so much the temporal nature of responses.

The domain of unsupervised cluster analyses constitutes another category of multivariate approaches. Unsupervised cluster analyses are a category of classification methods that are aimed to "identify structure in an unlabeled data set by objectively organizing data into homogenous groups where the within-group similarity is minimized and the between-group dissimilarity is maximized" (Liao, 2005). The advantage of unsupervised clustering approaches is that they do not require any normality or parametric assumptions within the data (Genolini and Falissard, 2010). In addition, in the case of longitudinal data, they do not require any assumptions with respect to the shape of the trajectory, meaning they can also handle non-linearity (Genolini and Falissard, 2010). As such, these unbiased and data-driven approaches analyses form an effective method to characterize individuals on a combination of traits.

Similar to PCA and factor analysis however, most clustering approaches are intended to cluster subjects based on static data (single measurements, (Liao, 2005)). In the field of human cohort studies, for example in epidemiology or clinical research, it is common to collect data on repeated time points and to record multiple variables per subject. One of the interests then often is to detect the presence of relatively homogenous subgroups of patients and delineate their characteristics (Subtil et al., 2017). When clustering longitudinal or repeated measures data it is important to account for the correlation between the different times that data was collected, and to adjust the distance measure used to partition the data, such that it is appropriate for time series (Liao, 2005). Furthermore, time series data are particularly sensitive to missing values (Molenberghs et al., 2004, Genolini, 2015). Genolini et al. (2015) therefore developed a clustering procedure that was specifically designed for concurrent clustering of multiple longitudinal response variables, or response trajectories, *kml3d*. This algorithm is an implementation of *k*-means clustering adapted to analysis of response trajectories. *K*-means is a partitional clustering algorithm that decomposes data into a set of *k* clusters through an iterative process that assigns subjects to its closest cluster center until there is a minimal decrease in

squared error (Frades and Mathiesen, 2010). The advantage of *kml3d* is that it allows for dependence between time points, and offers the selection between different distance measures that are suitable for time series analysis (Genolini et al., 2015). In addition, it provides several methods to deal with missing values, and produces a number of quality criteria to choose from when selecting the optimal number of clusters (Genolini et al., 2015). This particular clustering approach has been demonstrated to be of similar efficiency to latent class models (Genolini and Falissard, 2010). In the present thesis we therefore used the *kml3d* clustering to define mice on their individual response trajectories because it facilitated both our requirements.

Thesis outline

The aim of this thesis was twofold: to map inter-individual variability in behavioral and physiological responses in habituation of anxiety-related responses within and between multiple mouse inbred strains, by means of a multivariate, data-driven approach. And: to explore whether such identified inter-individual-variability in responses may be systematically employed in the design of animal experiments with the goal to improve the quality of experimental results. Throughout this thesis, we strived for research methods that would facilitate compliance with the 3R's and that would benefit the quality and generalizability (i.e. external validity) of our findings. For example, we used *in silico* methods in the form of retrospect analyses to answer our research questions where possible (**Chapters 2 and 5**), thereby reducing the number of animals needed to address our research objectives. All experimental chapters are reported in accordance with the Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines 2.0, which is a set of guidelines that was developed to improve and harmonize the reporting of animal studies, in order to benefit reproducibility in animal experimentation (Percie du Sert et al., 2020a; 2020b).

Furthermore, we systematically incorporated known sources of phenotypic variation in the design of our experimental work: we assessed inter-individuality in three mouse inbred strains commonly used in preclinical anxiety-research and known for their differential anxiety profiles: BALB/c, C57BL/6 and 129S2. In addition, we systematically varied the factor experimenter in the experimental design in **Chapters 3 and 4**, as suggested by Richter et al. (2020). To increase power, our experiments were designed and analyzed using factorial designs (**Chapters 2, 3 and 4**) and a randomized complete block design (**Chapter 4**), while all statistical inferential testing was conducted using generalized linear mixed models (**Chapters 2-5**).

Chapter 1 describes the main concepts used in this thesis and discusses these concepts in a more elaborate background. In **Chapter 2** we present an individual based, multidimensional approach to assess the change of anxiety-related and activity behavior over time. We used this approach on previously collected mHB-data to explore whether this would reveal subgroups of BALB/c and 129-mice that follow the same response over time in anxiety-related and activity behavior. We found that mice of various 129 sub-strains may differ in their ability to adapt to novelty. **Chapter 3** subsequently aimed to empirically validate the existence of subtypes of response in 129-mice and two additional, frequently used mouse inbred strains: BALB/c and C57BL/6. These strains are known for their differential anxiety profiles. In addition to individuality on a behavioral level, we measured blood plasma corticosterone concentrations to assess individual variability in corticosterone response. We found that individuals of all three strains indeed displayed subtypes of response, but did not extend these behavioral profiles to corticosterone responses.

With the empirical confirmation of subtypes of response in all three strains, we tested whether employing this information in the design of a pharmacological experiment would affect the quality of experimental results in **Chapter 4**. In a pre-experimental period, we characterized mice of the same three strains on their individual response type. With this information we designed a pharmacological experiment using a randomized complete block design in which we matched treatment-control pairs of mice on weight and individual response type in one half of the experimental pool, while mice were not matched on individual response type (weight only) in the other half of the experimental pool. We found that treatment, strain and experimenter effects differed between these two pools of mice, and concluded that incorporating individual variability in the design of an animal experiment may benefit the quality of experimental results. **Box 2** presents the between strain differences in behavior as measured in the mHB and corticosterone response that were recorded in the pre-experimental period of the study described in Chapter 4. **Chapter 5** explores the genetic underpinnings of habituation of anxiety responses by presenting a consomic strain survey that was conducted on previously collected mHB-data. This dataset consisted of a single five-minute mHB-trial, which for our purpose was transposed to five 60-second-epochs. The survey revealed a prominent role for mouse chromosome 19 in the change of anxiety related behavior over time in this CSS-panel. Mouse chromosome 19 indeed has been linked to several anxiety related behaviors, but as this study suggests it may also be linked to the change of anxiety related behavior over time. **Chapter 6** provides a general discussion of the main findings of all chapters and offers suggestion for further research.

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.

American Psychiatric Association 2013. Diagnostic and Statistical Manual of Mental Disorders. Fifth edition. American Psychiatric Association, Arlington, VA, USA.

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.12430https://doi.org/10.1111/2041-210X.12430.

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

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.

Baker, M. 2016. 1.500 scientists lift the lid on reproducibility. Nature 533, 452-454, https://doi.org/10.1038/533542a.

Bandelow, B., Michaelis, S. 2015. Epidemiology of anxiety disorders in the 21st centry. Dialogues Clin. Neurosci. 17 (3), 327-335, https://doi.org/10.31887/DCNS.2015.17.3/bbandelow.

Barnett, S.A. 1975. The Rat: a study in behavior. University of Chicago Press, Chicago, Il.

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

Baud, A., Flint, J. 2017. Identifying genes for neurobehavioral traits in rodents: progress and pitfalls. Dis. Models Mech. 10, 373-383, https://doi.org/10.1242/dmm.027789.

Beck, A.T., Emery, G., Greenberg, R.L. 1985. Anxiety disorders and phobias; A cognitive perspective. Basic Books: New York, NY.

Begley, C.G., Ioannidis, J.P.A. 2015. Reproducibility in science: improving the standard for basic and preclinical research. Circ. Res. 116 (1), 116-126, https://doi.org/10.1161/CIRCRESAHA.114.303819.

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, http://doi.org/10.3168/jds.2017-13978.

Belzung, C., Le Pape, G. 1994. Comparison of different behavioral test situations used in psychopharmacology for measurements of anxiety. Physiol. Behav. 56 (3), 623–628, https://doi.org/10.1016/0031-9384(94)90311-5.

Belzung, C., Berton, F. 1997. Further pharmacological validation of the BALB/c neophobia in the free exploratory paradigm as an animal model of anxiety. Behav. Pharmacol. 8 (6-7), 541–548, https://doi.org/10.1097/00008877-199711000-00012.

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

Belzung, C., Philippot, P. 2007. Anxiety from a phylogenetic perspective: is there a qualitative difference between human and animal anxiety? Neural Plast. 59676, https://doi.org/10.1155/2007/59676.

Beynen, A.C., Gärtner, K., van Zutphen, L.F.M. 2003. Chapter 5: Standardization of animal experimentation. In: Principles of Laboratory Animal Science, Revised edition. Elsevier, Amsterdam.

Blanchard, R.J., Yudko, E.B., Rodgers, R.J., Blanchard, D.C., 1993. Defense system psychopharmacology: an ethological approach to the pharmacology of fear and anxiety. Behav. Brain Res. 58, 155–165, https://doi.org/10.1016/0166-4328(93)90100-5.

Blanchard, D.C., Griebel, G., Blanchard, R.J. 2001. Mouse defensive behaviors: pharmacological and behavioral assays for anxiety and panic. Neurosci. Biobehav. Rev. 25 (3), 205-218, https://doi.org/10.1016/s0149-7634(01)00009-4.

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.

Boissy, A. 1995. Fear and fearfulness in animals. Q. Rev. Biol. 70, 165-191, https://doi.org/10.1086.418981.

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., Caldarone, B. J., Reilly, A. A., Flaherty, L., 2000. Habituation of activity in an open field: A survey of inbred strains and F1 hybrids. Behav. Genet. 30, 285-293, https://doi.org/10.1023/A:1026545316455.

Bothe, G.W.M., Bolivar, V.J., Vedder, M.J, Geistfeld, J.G., 2005. Behavioral differences among fourteen inbred mouse strains commonly used as disease models. Comp. Med. 55 (4), 326-334, PMID: 16158908.

Bourin, M., Petit-Demoulière, B., Nic Dhonnchadha, B., Hascöet, M. 2007. Animal models of anxiety in mice. Fundam. Clin. Pharmacol. 21(6), 567-574, https://doi.org/10.1111/j.1472-8206.2007.00526.x.

Bouwknecht, J. A., Paylor, R. 2008. Pitfalls in the interpretation of genetic and pharmacological effects on anxiety-like behavior in rodents. Behav. Pharmacol., 19(5-6), 385-402, https://doi.org/10.1097/FBP.0b013e32830c3658.

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.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376 https://doi.org/10.1038/nrn3475.

Canadian Council on Animal Care, 2019. Reproducibility is it a fourth R? https://ccac.ca/Documents/Publications/CCAC_Reproducibility-Is-it-a-Fourth-R.pdf.

Clément, Y., Calatayud, F., Belzung, C. 2002. Genetic basis of anxiety-like behaviour: A critical review. Brain. Res. Bull., 57 (1), 57-71, https://doi.org/S0361-9230(0)00637-2.

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

Cockrem, J. F., 2013. Individual variation in glucocorticoid stress responses in animals. Gen. Comp. Endocrinol. 181, 45-58, https://doi.org/10.1016/j.ygcen.2012.11.025.

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.

Crawley, J.N. 2000. What's wrong with my mouse? Behavioral Phenotyping of Transgenic and Knockout Mice, Second edition. Wiley-Liss, New-Jersey, USA.

Cryan, J.F., Holmes, A. 2005. The ascent of mouse: advances in modelling human depression and anxiety. Nat. Rev. Drug Discov. 4 (9), 775–790, https://doi.org/10.1038/nrd1825.

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.

Ducottet, C., Belzung, C., 2004. Behaviour in the Elevated-Plus-Maze predicts coping after subchronic mild stress in mice. Physiol. Behav. 81 (3), 417-426, https://doi.org/10.1016/j.physbeh.2004.01.013.

Dutta, S., Sengupta, P., 2016. Men and mice: relating their ages. Life Sci. 152, 244-248, http://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. 14, 65-64, https://doi.org/10.1016/j.cobeha.2016.11.016

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.

Ennaceur, A., Chazot, P.L. 2016. Preclinical animal anxiety research – flaws and prejudices. Pharmacol. Res. Perspect. 4(2), e00223, https://doi.org/10.1002/prp2.223.

Estanislau,C., 2012. Cues to the usefulness of grooming behavior in the evaluation of anxiety in the elevated plus-maze. Psychol. Neurosci. 5(1), 105-112, https://doi.org/10.3922/j.psns.2012.1.14.

Festing, M.F.W., 2004a. Refinement and reduction through the control of variation. Altern. Lab. Anim. 32 (1), 259-263, https://doi.org/10.1177/026119290403201s43.

Festing, M.F.W., 2004b. Good experimental design and statistics can save animals, but how can it be promoted? Altern. Lab. Anim. 32, 133-135, https://doi.org/10.1177/026119290403201s20.

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

Festing, M.F.W. 2016. Study Design, Ch3. In: Animal Models for Human Cancer: Discovery and Development of Novel Therapeutics. Wiley-VCH, Weinheim, https://doi.org/10.1002/9783527695881.ch3.

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

Feyissa, D.D., Aher, Y.D., Engidawork, E., Höger, H., Lubec, G., Korz, V. 2017. Individual differences in male rats in a behavioral test battery: A multivariate statistical approach. Front. Behav. Neurosci. 11, 26, https://doi.org/10.3389/fnbeh.2017.00026.

Frades, I., Matthiesen, R. 2010. Overview on techniques in cluster analysis. Methods Mol. Biol. 593, 81-107, https://doi.org/10.1007/978-1-60327-194-3_5.

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, http://doi.org/10.1126/science.1235294.

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, http://doi.org/10.3389/fvets.2018.00131.

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

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.

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

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/.

Guenet, J.L. 2005. The mouse genome. Genome Res. 15 (12), 1729-1740, https://doi.org/10.1101/gr.3728305.

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, http://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 behavior and risk assessment of inbred mice. Front. Behav. Neurosci. 8, 314, https://doi.org/10.3389/fn-beh.2014.00314.

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

Herman, J.P., McKlveen, J.M., Ghosal, S., Kopp, B., Wulsin, A., Makinson, R., Scheimann, J., Myers, B., 2016. Regulation of the hypothalamic-pituitary-adrenocortical stress response. Compr. Physiol. 6, 603-621, https://doi.org/10.1002/cphy.c150015.

Hettema, J.M., Neale, M.C., Kendler, K.S. 2001. A review and meta-analysis of the genetic epidemiology of anxiety disorders. Am. J. Psychiatry. 158(10), 1568–1578, https://doi.org/10.1176/appi.ajp.158.10.1568.

Hovatta, I., Tennant, R.S., Helton, R., Marr, R.A., Singer, O., Redwine, J.M., Ellison, J.A., Schadt, E.E., Verma, I.M., Lockhart, D.J., Barlow, C. 2005. Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature 438(7068), 662–666, https://doi.org/10.1038/nature04250.

Ioannidis, J.P.A. 2005. Why most published research findings are false. PLoS Med. 2, e124, https://doi.org/10.1371/journal.pmed.0020124.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang P. 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167 (7), 748–751, https://doi.org/10.1176/appi.ajp.2010.09091379.

International Committee on Standardized Genetic Nomenclature for Mice, Rat Genome and Nomenclature Committee. 2016. Guidelines for Nomenclature of Mouse and Rat Strains. http://www.informatics.jax.org/mgihome/nomen/strains.shtml

Irwin, J.R., McClelland, G.H., 2003. Negative consequences of dichotomizing continuous predictor variables. J. Mark. Res. 40 (3), 366-371, https://doi.org/10.1509/jmkr.40.3.366.19237.

Jakovcevski M., Schachner, M., Morellini, F. 2008. Individual variability in the stress response of C57BL/6J male mice correlates with trait anxiety. Genes Brain Behav. 7 (2), 235-243, https://doi.org/10.1111/j.1601-183X.2007.00345.x.

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

Kalueff, A.V., Tuohimaa, P. 2005. Mouse grooming microstructure is a reliable anxiety marker bidirectionally sensitive to GABAergic drugs. Eur J Pharmacol. 508(1-3), 147-153, https://doi.org/10.1016/j.ejphar.2004.11.054.

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.

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.

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

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.

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.

Korte, S.M., 2001. Corticosteroids in relation to fear, anxiety and psychopathology. Neurosci. Biobehav. Rev. 25 (2), 117-142, https://doi.org10.1016/S0149-7634(01)00002-1.

Kromer, S.A., Kessler, M.S., Milfay, D., Birg, I.N., Bunck, M., Czibere, L., Panhuysen, M., Pütz, B., Deussing, J.M., Holsboer, F., Landgraf, R., Turck, C.W. 2005. Identification of glyoxalase-I as a protein marker in a mouse model of extremes in trait anxiety. J. Neurosci. 25, 4375–4384, https://doi.org/10.1523/JNEUROSCI.0115-05.2005.

Laarakker, M.C., 2009. Hunting anxiety genes. A consomic survey to unravel the genetics of avoidance behavior in mice. Utrecht University, https://dspace.library.uu.nl/handle/1874/34756.

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, http://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, http://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, 95, https://doi.org/10.1186/s12863-016-0411-4.

Lander, E. S., Botstein, D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121 (1), 186-199, PMCID: PMC1203601.

Landgraf, R., Kessler, M.S., Bunck, M., Murgatroyd, C., Spengler, D., Zimbelmann, M., Nussbaumer, M., Czibere, L., Turck, C.W., Singewald, N., Rujescu, D., Frank, E., 2007. Candidate genes of anxiety-related behavior in HAB/LAB rats and mice: focus on vasopressin and gluoxalase-I. Neurosci. Biobehav. Rev. 31 (1), 89-102, https://doi.org/10.1016/jn.neubiorev.2006.07.003.

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

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

Liao, T.W. 2005. Clustering of time series data – a survey. Pattern Recognit. 38, 1857-1874, https://doi.org/10.1016/j.patcog.2005.01.025.

Liebsch, G., Montkowski, A., Holsboer, F., and Landgraf, R. 1998. Behavioural profiles of two Wistar rat lines selectively bred for high or low anxiety related behaviour. Behav. Brain Res. 94, 301–310, https://doi.org/10.1016/S0166-4328(97)00198-8.

Lister, R.G. 1987. The use of a plus-maze to measure anxiety in the mouse. Psychopharmacology 92, 180–185, https://doi.org/10.1007/BF00177912.

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

McEwen, B.S., 1999. Stress and hippocampal plasticity. Annu. Rev. Neurosci. 22, 105-122, https://doi.org/10.1146/annurev.neuro.22.1.105.

McNaughton, N., Corr, P.J., 2004. A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance. Neurosci. Biobehav. Rev. 28 (3), 285-305, https://doi.org/10.1016/j.neubiorev.2004.03.005.

Merikangas, K.R., Pine, D. 2002. American College of Neuropsychopharmacology. Genetic and Other Vulnerability Factors for Anxiety and Stress Disorders. In: Neuropsychopharmacology: The Fifth Generation of Progress. Lippincott, Williams & Wilkins; Nashville, TN, USA, p. 867.

Molenberghs. G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., Carroll, R.J. 2004. Analyzing incomplete longitudinal clinical trial data. Biostatistics, 5(3), 445–464, https://doi.org/10.1093/biostatistics/5.3.445.

Moy, S. S., Nadler, J. J., Young, N. B., Perez, A., Holloway, L. P., Barbaro, R. P., Barbaro, J. R., Wilson, L. M., Threadgill, D. W., Lauder, J. M., Magnuson, T. R., Crawley, J. N. 2007. Mouse behavioral tasks relevant to autism: phenotypes of 10 inbred strains. Behav. Brain Res., 6 (1), 4-20, https://doi.org/10.1016/j.bbr.2006.07.030.

Nadeau, J. H., Forejt, J., Takada, T., Shiroishi, T. 2012. Chromosome substitution strains: gene discovery, functional analysis, and systems studies. Mamm. Genome 23, 693-705, https://doi.org/10.1007/s00335-012-9426-y.

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

Ohl, F., Sillaber, I., Binder, E., Keck, M.E., Holsboer, F. 2001b. Differential analysis of behavior and diazepam-induced alterations in C57BL/6N and BALB/c mice using the modified hole board test. J. Psychiatr. Res. 35 (3), 147-154, https://doi.org/10.1016/S0022-3956(01)00017-6.

Ohl, F., 2003. Testing for anxiety. Clin. Neurosci. Res. 3 (4-5), 233-238, https://doi.org/10.1016/S1566-2772(03)00084-7.

Ohl,F. 2005. Animal models of anxiety. Handb Exp Pharmacol. 169, 35-69, https://doi.org/10.1007/3540-28082-0_2.

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.

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.

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lidster, K., MacCallum, C. J., Macleod, M., Petersen, O., Rawle, F., Reynolds, P., Rooney, K., Sena, E. S., Silberberg, S. D., Steckler, T., Würbel, H., 2020. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. BMC Vet Res. 16, 242, https://doi.org/10.1186/s12917-020-02451-y.

Percie du Sert N., Ahluwalia A., Alam S., Avey M. T., Baker M., Browne W. J., Clark A., Cuthill I. C., Dirnagl U., Emerson M., Garner P., Holgate S. T., Howells D. W., Hurst V., Karp N. A., Lazic S. E., Lidster K., MacCallum C. J., Macleod M., Pearl E. J., Petersen O. H., Rawle F., Reynolds P., Rooney K., Sena E. S., Silberberg S. D., Steckler T., Würbel H. 2020. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLos Biol. 18 (7), e3000411, https://doi.org/10.1371/journal.pbio.3000411.

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.

Réale, D., Dingemanse, N.J. 2012. Animal personality. eLS, https://doi.org/10.1002/9780470015902.aoo23570.

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., Garner, J.P., Auer, C., Kunert, J., Wuerbel, H. 2010. Systematic variation improves reproducibility of animal experiments. Nat. Meth. 7, 167-168, https://doi.org/10.1038/nmeth0310-167.

Richter, S.H. 2017. Systematic heterogenization for better reproducibility in animal experimentation. Lab Animal (NY) 46 (9), 343-349, https://doi.org/10.1038/laban.1330.

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

Rodgers, R.J., Cole, J.C., Cobain, M.R., Daly, P., Doran, P.J., Eells, J.R., Wallis, P. 1992. Anxiogenic-like effects of fluprazine and eltoprazine in the mouse elevated plus maze: profile comparison with 8-OH-DPAT, TFMPP and mCPP. Behav. Pharmacol. 3, 621–624, PMID: 11224163.

Rodgers, R.J., Cao, B-J., Dalvi, A., Holmes, A. 1995. Animal models of anxiety: an ethological perspective. Braz. J. Med. Biol. Res. 30, 289–304, https://doi.org/10.1590/s0100-879x1997000300002.

Rodgers, R.J., Johnson, N.J., 1995. Factor analysis of spatiotemporal and ethological measures in the murine elevated plus-maze test of anxiety. Pharmacol. Biochem. Behav. 52 (2), 297–303, https://doi.org/10.1016/0081-3057(95)00138-m.

Rodgers, R., Haller, J., Holmes, A., Halasz, J., Walton, T., Brain, P., 1999. Corticosterone response to the plus-maze: high correlation with risk assessment in rats and mice. Physiol. Behav. 68 (1), 47-53, https://doi.org/10.1016/S0031-9384(99)00140-7.

Rougé-Pont, F., Deroche, V., Le Moal, M., Piazza, P. V., 1998. Individual differences in stress-induced dopamine release in the nucleus accumbens are influenced by corticosterone. Eur. J. Neurosci. 10 (12), 3903-3907, https://doi.org/10.1046/j.1460-9568.1998.00438.x.

Russel, W.M.S., Burch, R.L., 1959. The principles of humane experimental technique. Methuen & Co., London, UK.

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, http://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, http://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, http://doi.org/10.1111/j.1601-183X.2009.00527.x

Salomons, A.R., Espititia Pinzon, N., Boleij, H., Kirchhoff, S., Arndt, S.S., Nordquist, R.E., Lindemann, L., Jaeschke, J., Spooren, W., Ohl, F. 2012. Differential effects of diazepam and MPEP on habituation and neuro-behavioral processes in inbred mice. Behav. Brain Funct. 8, 30, https://doi.org/10.1186/1744-9081-8-30.

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, http://doi.org/10.1016/j.bbr.2013.02.023.

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

Sgoifo, A., de Boer, S. F., Haller, J., Koolhaas, J. M., 1996. Individual differences in plasma catecholamine and corticosterone stress responses of wild-type rats. Physiol. Behav. 60 (6), 1403-1407, https://doi.org/10.1016/S0031-9384(96)00229-6.

Singer, J. B., Hill, A. E., Nadeau, J. H., Lander, E. S. 2005. Mapping quantitative trait loci for anxiety in chromosome substitution strains of mice. Genetics, 169 (2), 855-862, https://doi.org/10.1534/genetics.104.031492.

Song, M.K., Lin, F.C., Ward ,S.E., Fine ,J.P. 2013. Composite variables: when and how. Nurs. Res. 62 (1), 45-49, https://doi.org/10.1097/NNR.0b013e3182741948.

Stegman, Y., Schiele, M.A., Schümann, D., Lonsdorf, T.B., Zwanzger, P., Romanos, M., Reif, A., Domschke, K., Deckert, J., Gamer, M., Pauli, P. 2019. Individual differences in human fear generalization – pattern identification and implications for anxiety disorders. Transl. Psychiatry, 9, 307, https://doi.org/10.1038/s41398-019-0646-8.

Taft, R.A., Davisson, M., Wiles, M.V., 2006. Know thy mouse. Trends Genet. 22 (12), 649-653, https://doi.org/10.1016/j.tig.2006.09.010.

Tam, W. Y., Cheung, K-K., 2020. Phenotypic characteristics of commonly used mouse inbred strains. J. Mol. Med. Jul 25. https://doi.org/10.1007/s00109-020--1953-4.

Taylor, J.M., Whalen, P.J. 2015. Neuroimaging and anxiety: the neural substrates of pathological and non-pathological anxiety. Curr. Psychiatry Rep. 17, 49, https://doi.org/10.1007/s11920-015-0586-9.

Tebbich, S., Fessl, B., Blomqvist, D. 2009. Exploration and ecology in Darwin's finches. Evol. Ecol. 23, 591-605, https://doi.org/10.1007/s10682-008-9257-1.

Trullas, R., Skolnick, P. 1993. Differences in fear motivated behaviors among inbred mouse strains. Psychopharmacology, 111 (3), 323 – 331, https://doi.org/10.1007/BF02244948.

Turri, M.G., Datta, S.R., DeFries, J., Henderson, N.D., Flint, J. 2001. QTL analysis identifies multiple behavioral dimensions in ethological tests of anxiety in laboratory mice. Curr Biol. 11(10), 725-734, https://doi.org/10.1016/s0960-9822(01)00206-8.

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, http://doi.org/10.1038/s41592-018-0224-7.

Van der Staay, F.J. 2006. Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. Brain Res. Rev. 52, 131-159, https://doi.org/10.1016/j.brainresrev.2006.01.006.

Van der Staay, F.J., Arndt, S.S., Nordquist, R.E. 2009. Evaluation of animal models of neurobehavioral disorders. Behav. Brain Funct. 5, 11, https://doi.org/10.1186/1744-9081-5-11.

van der Staay, F. J., Arndt, S. S. & Nordquist, R. E. 2010. The standardization–generalization dilemma: a way out. Genes, Brain Behav. 9, 849–855, https://doi.org/10.1111/j.1601-183X.2010.00628.x.

van der Staay, F.J., Steckler, T. 2002. The fallacy of behavioral phenotyping without standardization. Genes Brain Behav. 1 (1), 9-13, https://doi.org/10.1046/j.1601-1848.2001.00007.x.

Vásquez, R. 1994. Assessment of predation risk via illumination level: facultative central place foraging in the cricetid rodent *Phyllotis darwini*. Behav. Ecol. Sociobiol. 34, 375-381.

Vidal-Gómez, J. 2016. Consistent individual differences in some behaviors in mice of the C57BL/6J strain. Anu. De Psicol. 46 (2), 83-89, https://doi.org/10.1016/j.anpsic.2016.07.005.

Voelkl, B., Würbel, H., 2019. A Reaction Norm Perspective on Reproducibility. bioRxiv, 510941, https://doi.org/10.1101/510941.

Voelkl, B., Altman, N.S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N.A., Kas, M.J., Schielzeth, H., Van de Casteele, T., Würbel, H. 2020. Reproducibility of animal research in the light of biological variation. Nat. Rev. Neurosci. 2020; 21, 384-393, https://doi.org/10.1038/s41583-020-0313-3.

Voelkl, B., Würbel, H., Krzywinski, M., Altman, N. 2021. The standardization fallacy. Nat. Meth. 18, 5-7, https://doi.org/10.1038/s41592-020-01036-9.

Wahlsten, D. 2011. Mouse Behavioral Testing: How to use mice in behavioral neuroscience, Chapter 2. Elsevier Academic Press.

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

Wermter A.K., Laucht, M., Schimmelmann, B.G., Banaschewski, T., Sonuga-Barke, E.J.S., Rietschel, M., Becker, K. 2010. From nature versus nurture, via nature and nurture, to gene x environment interaction in mental disorders. Eur. Child Adolesc. Psychiatry 19(3), 199–210, https://doi.org/10.1007/s00787-009-0082-z.

Würbel, H. 2000. Behaviour and the standardization fallacy. Nat. Genet. 26, 263, https://doi.org/10.1038/81541.

Young, E.A. Abelson, J.L. Liberzon, I. 2008. Stress Hormones and Anxiety Disorders. In: Handbook of Anxiety and Fear. Academic Press, Oxford, p. 455-473.

Zocher, S., Schilling, S., Grzyb, A.N., Adusumilli, V.S., Bogado Lopes, J., Günther, S., Overall, R.W., Winter, Y., Kempermann, G., 2020. Early-life environmental enrichment generates persistent individualized behavior in mice. Sci. Adv. 6 (35), eabb1478, https://doi.org/10.1126/sciadv.abb1478.

Chapter 2

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

Journal of Neuroscience Methods, 2020, 343, 108810

Marloes H. van der Goot[1, 3], Hetty Boleij[1], Jan van den Broek[2], Amber R. Salomons[1], Saskia S. Arndt[1], Hein A. van Lith[1,3]

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

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.

Keywords
Phenotypic variation, inbred mice, behavioral profile, habituation, sensitization, cluster analysis

Abbreviations
mHB: modified Hole Board, CVI: Clustering Validity Index; GLMM: Generalized linear mixed models

Declaration of interest
Declaration of interest: none

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

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 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; Casarrubea et al. 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; Salomons et al. 2010b; Salomons et al. 2010c; Salomons et al. 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; Laarakker et al. 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; Salomons et al. 2010b; Salomons et al. 2010c; Salomons et al. 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. AppendixTable 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), 129X1/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 x 215 x 141 mm, floor area 370 cm^2) or Macrolon Type II L cages (size: 365 x 207 x 140 mm, floor area 530 cm^2, Techniplast, Milan, Italy) with standard bedding material (autoclaved Aspen Chips, Abedd-Dominik Mayr KEG, Köflach, Austria) and a tissue (KLEENEX® Facial Tissue, Kimberley-Clark Professional BV, Ede, the Netherlands) and 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°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 x 50 x 50 cm) with a board made of the same material (60 x 20 x 20 cm) functioning as an unprotected area, as it is positioned in the center of box. The board stacks 20 cylinders (diameter 15 mm) in three lines (Figure 1). The area around the board is divided into 10 rectangles (20 x 15 cm) and 2 squares (20 x 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 lux or 120 lux, depending on the study (see appendix Table A1).

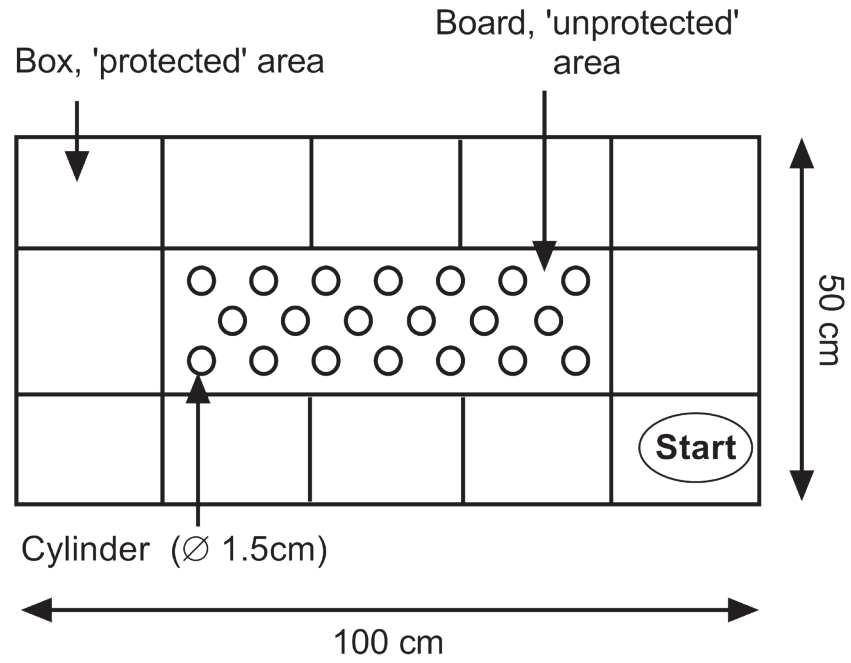


Figure 1. Schematic overview of the modified Hole Board.

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 minutes, 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 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; Laarakker et al. 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; Salomons et al. 2010b; Salomons et al. 2010c; Salomons et al. 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

$$Z_T = \frac{X_T - \mu_T}{\sigma_T}; Z_L = \frac{X_L - \mu_L}{\sigma_L}; Z_R = \frac{Z_T + (-Z_L)}{2}$$

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 Fallissard, 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):

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: $\alpha = 1-[1-0.05]^{1/\lambda,}$ where λ = 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 Figure 1. The differences between clusters in residual integrated behavioral z-scores on trial 1 were graphed as means with 95% confidence intervals in Figure 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 (Figure 3).

2.7 Ethical note

All experimental protocols were approved by the Animal Experiments Committee of the Academic Biomedical Center Utrecht, 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 Figure 2 (presented as mean residual integrated behavioral z-scores with 95% CI). Figure 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 Figure 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).

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.

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 Figure 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.

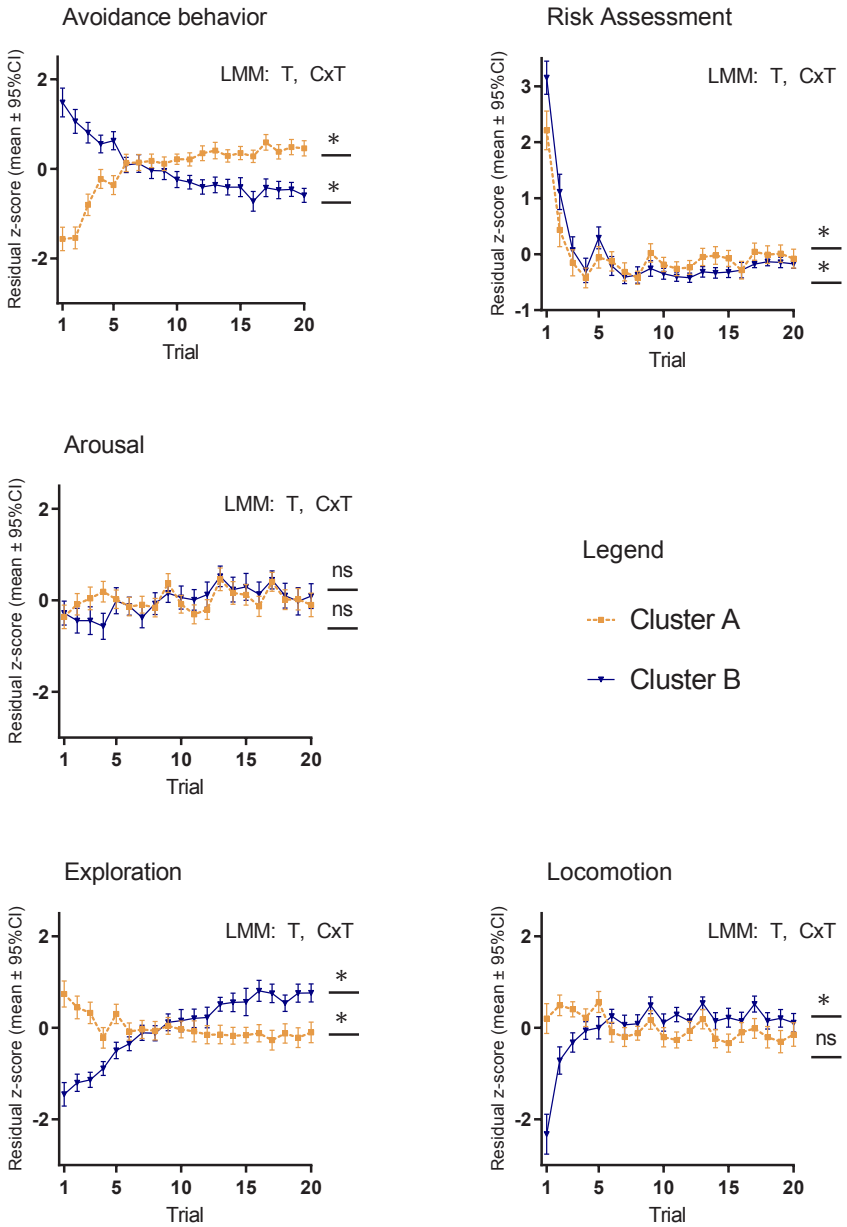


Figure 2. Residual integrated behavioral z-scores for mice in clusters A and B. Results are presented as means with 95% CI. Effects were significant in the linear mixed models (LMM) when $P < 0.05$. T indicates a significant main effect of Trial; CxT indicates a significant interaction between cluster 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.

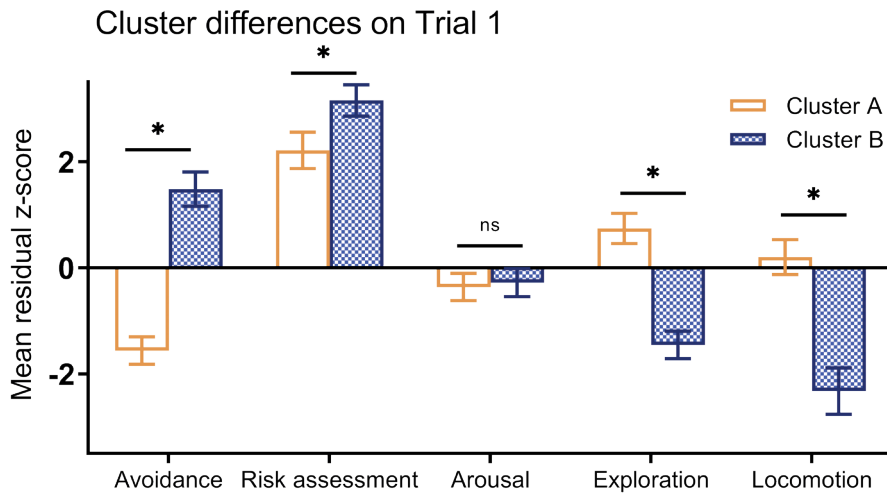


Figure 3. Initial levels on al behavioral dimensions for clusters A and B (mean residual z-score on trial 1). Results are summarized as estimated marginal means with 95% confidence intervals. * = Significant difference ($P < 0.05$) in *post hoc* comparison, ns = difference not significant.

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 Figure 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 Figure 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 Figure 2. Visual inspection of the data (Figure 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 Figure 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 exploration on trial 1 (2.194 ± 0.146, $t(123) = 15.039$ $P < 0.0001$, $d = 2.751$), see Figure 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 Figure 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, $t(2336) = 1.794$, $P = 0.0729$), 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 ($d = 1.827$), Figure 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'.

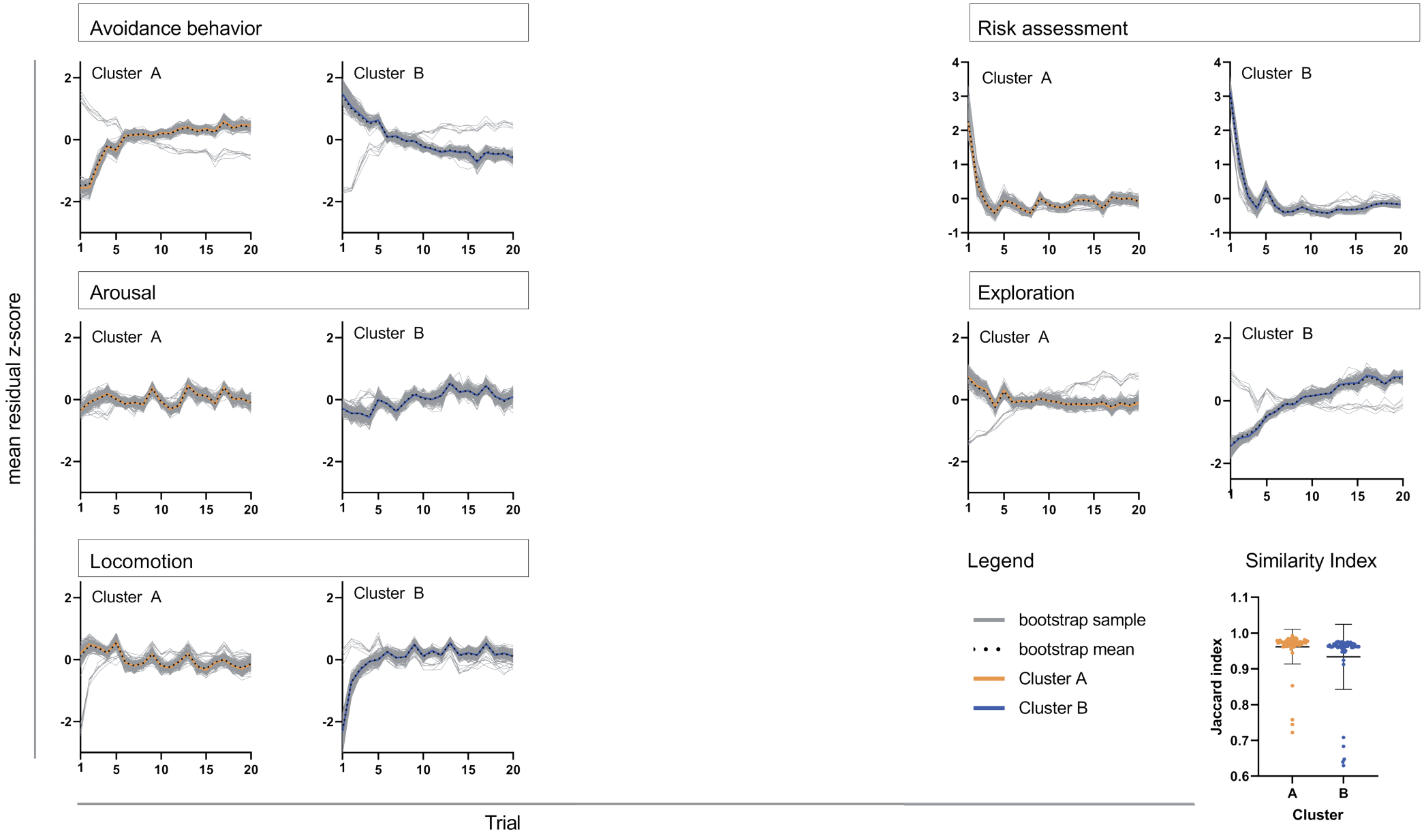


Figure 4. Mean trajectories (residual z-scores) of 200 bootstrap samples (grey) for each cluster, on each behavioral dimension. For visual comparison, the average trajectory of all bootstrap samples is depicted (black dotted line) against the trajectory belonging to cluster A (orange) or B (blue). These 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).

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. Figure 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. Figure 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.

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.

4. Discussion

The current paper explored inter-individual variability in habituation and sensitization responses in two mouse inbred strains. We re-inspected 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; Salomons et al. 2010b; Salomons et al. 2010c; Salomons et al. 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 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 129J 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; Salomons et al. 2010b; Salomons et al. 2010c; Salomons et al. 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 (Figure 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.

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.

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. 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* 1986; 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. http://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. http://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, http://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, http://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 (reprinted 2005, 2006). Chapter 5 – Standardization of animal experimentation. *In*: Principles of Laboratory Animal science. A contribution to the humane use and care of animals and to the quality of experimental results, 2nd edition (eds., van Zutphen, L.F.M., Baumans, V., Beynen, A.C.). Elsevier Science Publishers, Amsterdam, The Netherlands, pp. 103110.

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, http://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, http://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, http://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, http://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, http://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. 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, http://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. 14, 65-64, https://doi.org/10.1016/j.cobeha.2016.11.016.

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, http://doi.org/10.1016/j.physbeh.2013.06.007.

Festing, M., 2011. Inbred strains and toxicity testing. In: Proceedings of the Eleventh FELASA Symposium and the 40th Scand-LAS symposium – New Paradigms in Laboratory Animal Science (ed. Kaliste, E.), 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 Journal 55 (3), 399-404, http://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, http://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, http://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 Journal 46 (2), 106-117, http://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, http://doi.org/10.1093/ije/dyr219. Reprint of Lab. Anim. 1990; 24 (1), 71-77, http://doi.org/10.1258/002367790780890347.

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/.

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.

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, http://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., 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, http://doi.org/10.1016/j.yfrne.2010.04.001.

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.

Kryszczuk, K., Hurley, P., 2010. Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. In: El Gayar, N., Kittler, J., Roli, F. (eds). Multiple Classifier Systems. MCS 2010. Lecture Notes in Computer Science, vol 5997. Springer, Berlin, Heidelberg. 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, http://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, http://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 Genetics 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, http://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. Guidelines for the Care and Use of Mammals in Neuroscience and Behavioral Research. Washington, National Academies Press; 2003.

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, http://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 Instr. 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. Lab Animal (NY) 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. Third edition. New York: McGraw Hill, p. 699, 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, http://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, http://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, http://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, http://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 234, 2-1, https://doi.org/10.1016/j.jneumeth.2014.03.001.

Staay, van der F.J., Arndt, S.S., Nordquist, R.E., 2010. The standardization-generalization dilemma: a way out. Genes Brain Behav. 9, 849-855, https://doi.org/10.1111/j.1601-183X.2010.00628.x

Tibshirani, R., Walther, G., Hastie, T., 2002. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. 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, http://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. New York, NY: Springer, https://doi.org/10.1007/978-0-387-87458-6.

Appendix A

Table A1. Overview of factors that varied between experiments.

Study	(Sub) strains	Supplier[1]	Sex	Age at test (wks)	Housing location[2]	Experimenter	Cage size	Light on board	Observer version	Approval no
Salomons et al. (2010[a])	BALB/cJ (N=10) 129P3/J (N=10)	J	F	~8	NVI	D	Eurostandard Type II	120 lux	5.0	2007.I.01.007
Salomons et al. (2010[b])	129P3/J (N=13)	J	M	~9	UU-P	C	Eurostandard Type II L	120 lux	4.0	2007.I.01.007
Salomons et al. (2010[c])	I. BALB/cJ (N=8) 129P3/J (N=8)	J	M	~9	UU-P	A	Eurostandard Type II L	50 lux	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 lux	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 lux	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 lux	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 lux	5.0	2009.I.10.079

[1] 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.

[2] 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[1]
Anxiety related behavior		
- Avoidance behavior	Total number of board entries	-z
	Latency until first board entry	z
	Percentage of time spent on the board	-z
- Risk assessment	Total number of stretched attends	z
	Latency until first stretched attend	-z
- Arousal	Total number of self-groomings	z
	Latency until the first self-grooming	-z
	Percentage of time self-grooming	z
	Total number of boli	z
	Latency until first boli is produced	-z
Activity		
- Exploration	Total number of rearings in the box	z
	Latency until first rearing in the box	-z
	Total number of rearings on the board	z
	Latency until first rearing on the board	-z
	Total number of hole explorations	z
	Latency until first hole exploration	-z
	Total number of hole visits	z
	Latency until first hole visit	-z
- Locomotion	Total number of line crossings	z
	Latency until first line crossing	-z

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

Chapter 3

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

Physiology and Behavior, 2021, 239, 113503.

Marloes H. van der Goot[1,2], Melissa Keijsper[1], Annemarie Baars[1], Lisa Drost[1], Judith Hendriks[1], Susanne Kirchhoff[1], José G. Lozeman-van 't Klooster[1], Hein A. van Lith[1,2], Saskia S. Arndt[1]

[1] Section Animals in Science and Society, Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands

[2] Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands

Abstract

Inter-individual variability in behavioral and physiological response has become a well-established phenomenon in animal models of anxiety and other disorders. Such variability is even demonstrated within mouse inbred strains. A recent study showed that adaptive and non-adaptive anxiety phenotypes (measured as habituation and/or sensitization of anxiety responses) may differ within cohorts of 129 mice. This variability was expressed across both anxiety- and activity-related behavioral dimensions. These findings were based however on re-analysis of previously published data. The present study therefore aimed to empirically validate these findings in 129 mice. In addition, we assessed such inter-individuality in two other strains: BALB/c and C57BL/6. Males of three mouse inbred strains (BALB/c, C57BL/6 and 129S2) were behaviorally characterized through repeated exposure to a mild aversive stimulus (modified Hole Board, 4 consecutive trials). Behavioral observations were supplemented with assessment of circulating corticosterone levels. Clustering the individual response trajectories of behavioral and endocrine responses yielded two multidimensional response types of different adaptive value. Interestingly, these response types were displayed by individuals of all three strains. The response types differed significantly on anxiety and activity related behavioral dimensions but not on corticosterone concentrations. This study empirically confirms that adaptive capacities may differ within 129 cohorts. In addition, it extends this inter-individual variability in behavioral profiles to BALB/c and C57BL/6. Whether these two sub-types constitute differential anxiety phenotypes may differ per strain and requires further study.

Keywords
inter-individual variability, inbred mice, anxiety, habituation, cluster analysis

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

1. Introduction

Inter-individual variability in emotional reactivity to environmental challenges is a well-established phenomenon in animal models of stress, anxiety, depression, and post-traumatic stress disorder (e.g. [1][2][3][4][5][6]). This type of variability has repeatedly been associated with complex interactions between genetic and environmental factors, which are partly modulated by epigenetic processes [7][8]. Inter-individual variability has become of increasing interest in animal models that study the underlying mechanisms and/or treatment of psychiatric diseases [3][6]. In humans, the susceptibility to develop psychopathologies and the response to treatment is known to vary greatly between patients [3]. Similar circumstances may trigger development of affective disorders in some, while other individuals are unaffected [9]. Incorporating this variation in animal models may therefore not only make these models more representative [6], but could also improve our understanding of the underlying mechanisms that are involved in this differential susceptibility [3][9]. It has been suggested that a starting point for mapping such differential susceptibility could be a more in depth, individual based, characterization of behavior of a particular animal model [3]. This could then be followed by the identification of biological markers that may explain the differences between these sub-groups [3][9].

Behavioral habituation and sensitization are two contrasting forms of learning, and are defined as either the decremental (habituation) or incremental (sensitization) change in behavioral response after repeated exposure to environmental stimuli (provided these stimuli are not accompanied by biologically significant consequences)[10]. In rodents, exposure to novelty induces a biologically adaptive anxiety response that enables individuals to respond appropriately to potential threat [11]. In an adaptive phenotype, repeated exposure to such stimuli results in habituation (i.e. the waning) of anxiety-related behavior, enabling individuals to adapt to environmental challenges [11][14]. Several studies suggest that the opposite of a habituation response (i.e. a sensitization of anxiety behavior) may reflect a non-adaptive anxiety phenotype, that can be used as an indicator of pathological anxiety in rodent models [11][13][14][15][16]. In these studies, two strains that differed in innate emotionality [17] were repeatedly exposed to a mild aversive stimulus (the modified Hole Board). BALB/cJ (known as a neophobic mouse strain [18]) displayed initial high levels of anxiety that decreased with repeated exposure to the test. At the same time, exploration and locomotion increased, which, taken together reflected successful habituation. In contrast, 129 mice consistently

showed low initial levels of anxiety-related behavior that increased with repeated exposure, suggesting a sensitizing anxiety response [11][13][14][15][16]. This non-adaptive behavioral profile was further characterized by a lower expression of the immediate early gene c-Fos (a marker for neural activity) in the prelimbic cortex and lateral septum, brain areas involved in integration of emotional and cognitive processes, compared to rapidly habituation BALB/c.

Re-inspection of the behavioral data from these studies however demonstrated that responses may differ between individuals within these strains [19]. Using a multivariate cluster analysis on the combined data from these studies, van der Goot et al. [19] identified two homogenous subgroups of mice that followed the same response across trials: a habituation and a sensitization cluster. These clusters were found to be multidimensional, with individual mice consistently grouping together across dimensions indicative of anxiety related behavior, but also activity behavior [19]. The profiles of these subtypes mirrored the BALB/c specific habituation response, and the sensitization response that was characteristic for 129 mice. These individual based analyses however also revealed that a sub-population of 129 mice displayed a successful habituation response, which was overlooked when only comparing average strain responses. These identified behavioral profiles however were based on retrospect analyses on a dataset that consisted of multiple experiments. These studies were conducted over a time span of 4 years and varied naturally in experimenter, test location, time of year etcetera – all factors that are known to affect variability between experiments [20][21]. Therefore, the question remained to what extent the observed inter-individual variability, and its expression within and across strains was representative of variation in (sub-) types of response BALB/c and 129 mice in general, or whether the identified clusters were part result of the mere variation that was inherent to analyzing a dataset consisting of multiple experiments.

The goal of the present experiment therefore was to empirically validate the previous findings, and assess inter-individuality in adaptive capacities in a controlled experiment. Three mouse inbred strains were behaviorally characterized by repeated exposure to the modified Hole Board. Two strains - BALB/c and 129S2 - were also observed in the previous studies [13][14][15][16]. In addition, we assessed inter-individual variability in habituation and sensitization responses in an additional strain: C57BL/6. According to the data presented by The Jackson Laboratory these three inbred lines are the most frequently used mouse strains in biomedical research. Phenotypic characteristics of these strains have been reviewed elsewhere [22]. C57BL/6 mice are typically classified as non-anxious and highly active [23][24][25][26]. When repeatedly exposed to the modified Hole Board, these mice are characterized as highly active, displaying low levels of anxiety-related behavior and showing no further habituation to the test [27]. Mice were individually characterized on the same five behavioral dimensions that comprised the previously identified clusters [19]: avoidance behavior, risk assessment, arousal, exploration and locomotion.

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

2. Materials and Methods

2.1. Ethical statement

The experimental protocol was approved by the Central Animal Experiments Committee, The Hague, the Netherlands (CCD approval numbers: AVD1080020172264 and AVD1080020172264-1). The decision for approval was based on the Dutch implementation of EU directive 2010/63/EU (Directive on the Protection of Animals Used for Scientific Purposes). The experiment was conducted according to the Dutch 'Code on Laboratory Animal Care and Welfare'. Furthermore, the present animal study is reported to the best of our abilities according to the revised ARRIVE guidelines (ARRIVE 2.0; https://www.nc3rs.org.uk/revision-arrive-guidelines [35][36].

2.2. Animals and housing

This study tested naïve males of three mouse inbred strains: BALB/cAnNCrl (hereafter C, see http://www.informatics.jax.org/mgihome/nomen/strains.shtml#labcodes, *n* = 40, albino), C57BL/6NCrl (B6N, *n* = 40, black) and 129S2/SvPasCrl (129S2, *n* = 38, agouti). An additional two 129S2 mice died due to health reasons unrelated to the study and were not tested. The sample size

was determined beforehand and based on recommendations for cluster analysis by Dolnicar et al. [37]. This study assessed inter-individuality in male mice only. It has reliably been established that incorporating both sexes when analyzing between group factor (strain, sex etc.) and/or treatment effects in factorial designs does not necessitate a duplication of sample size [39][40][41]. To our knowledge however, it is unclear whether the same mechanism applies to unsupervised clustering techniques, such as the one used in this study. Unsupervised clustering approaches do not have any a priori assumptions regarding the distribution and variance of the data [40]. This unsupervised nature makes these analyses more sensitive to anomalies in the data than classical statistical approaches, and it has been established that the detection of meaningful clusters increases with sample size [37]. To ensure a maximal sample size, while maintaining a manageable technical load due to repeated testing in multiple inbred strains, it was therefore decided to first explore this variability in males, with the intention to extend potential findings to females in a follow up study.

Animals were bred by and purchased from Charles River Germany (Sulzfeld, Germany) and arrived at the research facility in four batches of $n = 10$ per strain. All mice were 6-8 weeks old upon arrival (mean body weight ± SEM and range of strains per batch are in supplementary Table S1). All animals were housed at the Central Laboratory Animal Research Facility of Utrecht University. Testing took place in the same rooms as where the animals were housed, and test equipment was placed in each room prior to arrival of the animals. Mice were housed individually in Macrolon Type II L cages (size: 365 x 207 x 140mm, floor area 530cm^2, Techniplast, Milan, Italy) with standard bedding material (autoclaved Aspen Chips, Abedd-Dominik Mayr KEG, Köflach, Austria) and a tissue (KLEENEX® Facial Tissue, Kimberley-Clark Professional BV, Ede, the Netherlands) and a PVC-shelter (Plexx BV, Elst, the Netherlands) as enrichment. Our previous research demonstrated that stress-levels in individually housed male mice did not differ significantly from socially housed male mice [38]. Food (CRM, Expanded, Special Diets Services Witham, UK) and tap water were available *ad libitum*.

Upon arrival mice were randomly allocated to one of two laboratory animal housing rooms for a habituation period of 17 days under a reversed 12h light/12h dark cycle (lights off at 7:00 AM) with a radio playing constantly as background noise. The number of mice per strain was kept similar between testing rooms. Relative humidity (mean percentage ± SD) was controlled (room A: 61.8% ± 3.93, range 48.3% – 81.0%; room B: 62.4% ± 4.16, range 50.5% – 81.1%) with a ventilation rate of 15-20 changes/hour and an average room temperature (mean ^{o}C ± SD) of 22.1^{o}C ± 0.33 (range 20.4 – 23.6) and 22.2^{o}C ± 0.33 (range 20.3 – 23.7) for rooms A and B respectively. Both animal rooms also housed female C57BL/6NCrl mice throughout the entire duration of the study. The mice were handled three times a week during the habituation period by the same experimenters that conducted the behavioral observations. During handling mice were accustomed to a polycarbonate clear mouse handling tube (Plexx BV, Elst, the Netherlands) according to the protocol from Gouveia & Hurst [43]. One handling tube was used per room to habituate the mice. The tube was cleaned with water and a damp tissue between animals. In addition, mice were picked up at the tail base to habituate them to the blood collection procedure for corticosterone measurements (see *2.4 Experimental protocol and blood sampling*).

2.3. Modified Hole Board

Mice were tested in the modified Hole Board (mHB), a test for assessment of unconditioned behavior that combines characteristics of an open field, a hole board and a light-dark box [44][45]. The mHB allows for analyzing a range of anxiety and activity related behaviors and as such is suitable for a complete phenotyping of complex behavioral constructs, such as behavioral habituation of anxiety responses [44]. The paradigm has been described extensively elsewhere [45] and is only briefly explained here. The apparatus consists of a grey PVC opaque box (100 x 50 x 50 cm) with a board made of the same material (60 x 20 x 20 cm) functioning as an unprotected area, as it is positioned in the center of box. The board stacks 20 cylinders (diameter 15 mm) in three lines. The area around the board is divided into 10 rectangles (20 x 15 cm) and 2 squares (20 x 20 cm). The periphery was illuminated with red light (1-5 lux) and functioned as the protected area. In contrast, the central board was illuminated by an additional stage light (120 lux) in order to increase the aversive nature of the central (unprotected) area.

2.4. Experimental protocol and blood sampling

Mice were behaviorally characterized through repeated exposure to the mHB. Testing occurred between 09:30 AM and 2 PM, during the active phase of the animals. Each batch was tested on 4 consecutive testing days within a single week. All mice were tested individually for a total of 4 consecutive trials. Each trial lasted 5 min and test order within batch was randomized across strains. At the start of the first trial, mice were transferred from the home cage to the mHB using the handling tube, and always placed in the same corner, facing the

central board. During the test, mice were allowed to freely explore the mHB-set up. Between trials mice were transported back to their home cage using the handling tube and the mHB was carefully cleaned with water and a damp towel before the next trial commenced. Behavior was scored live using the software Observer version 12.5 (Noldus Technology, Wageningen, the Netherlands). In addition, trials were recorded on video camera for raw data storage. Behavioral observations were conducted by two trained observers, each of which always tested in the same housing room. Inter-observer reliability was established prior to the start of the study at a moderate to good level [46] with an average Cohen's κ = 0.74 (range 0.67 - 0.84) over an average percentage agreement of 94% (range 89.6 – 97.14). Intra-observer reliability was established at a good level for both experimenters (Experimenter I, average Cohen's κ = 0.83, range 0.78 - 0.87; Experimenter II: average Cohen's κ = 0.85, range 0.70 – 0.95) over average percentage agreements of 93.35 % (range 92.18-95.19) and 97.28% (range = 95.94 – 98.6) respectively.

Circulating plasma corticosterone levels (pCORT) were assessed for each individual at three sampling moments. The first blood sample was taken one week prior to the behavioral test (7 days ± 1). The second sample was taken directly after behavioral testing, approximately 30 min after the first mHB trial. The third sample was collected one week after behavioral testing (7 days ± 1). For each individual mouse, it was ensured that all three samples were collected on approximately the same time of day to avoid fluctuation of pCORT due to circadian rhythm [47]. In order to determine baseline pCORT levels in rodents, for the first and the last blood sample, the average time from picking up the home cage from the home shelf to finishing the blood collection was recorded (first sample: 120.8 ± 43.4, range 58 – 319; third sample: 127.1 ± 33, range 61 - 276, time in seconds). With an average collection time of 120s pCORT levels on both levels were considered baseline/not-affected by handling stress. Blood sampling was conducted dropwise via tail vein incision [49], using a single edge industrial blade (GEM®: SPI Supplies, West Chester, PA, USA) by experienced technicians who were not involved in behavioral observations. Sampling occurred in a separate room, mice were transported to this location in their home cage, which was covered with a blanket because the corridor between the two rooms was not under a reversed light-regime. Blood drops were collected in pre-chilled EDTA coated Microvette® CB300 capillaries (Sarstedt, Nümbrecht, Germany) and stored on ice until plasma was collected by centrifuging the capillary tubes for 30 min on 4 ^{0}C and 3000 rpm (diameter of the rotor: 17 cm) in a centrifuge (IEC Microlite/Microlite RF®: Thermo Electron Cooperation; West Sussex, UK). Next, plasma (10-20 µl) was pipetted into microtubes and stored at -26 ^{0}C until further analysis.

2.5. Corticosterone response

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

2.6. Behavioral variables

Behavioral patterns were assessed by scoring behaviors listed in supplementary Table S2. These behaviors were scored as separate variables during testing. However, previous research has shown that the separate behavioral variables scored in the mHB can be reliably allocated to five behavioral dimensions: avoidance behavior, arousal, risk assessment, exploration and locomotion [50][51][52]. In a previous study the identified clusters were composed of these five dimensions [19]. Therefore these same dimensions were again included in the present study. The separate variables were summarized to their corresponding dimension by using the method of integrated behavioral z-scoring. This method was first proposed by Guilloux et al. [53] and further extended by Labots et al. [52] as a method for behavioral phenotyping in mice. The exact procedure is described in detail elsewhere [19][52] and will therefore only be described briefly here. In short, behavioral variables that measured different aspects (or different units) of the same behavioral dimension were normalized and combined to a single score representing that particular behavioral dimension or motivational system. Normalization was done by z-score transformation, which measures the amount of standard deviations each observation is above or below the mean of a reference group [52]. The transformed separate variables were averaged within each behavioral dimension. In the present experiment we used the pooled data (across all strains) as the reference group, as suggested by Labots et al. [52]. Supplementary Table S2 presents an overview of all included variables, per behavioral dimension.

2.7. Statistical analyses

2.7.1. Missing values and outliers

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

2.7.2. Analysis general

All analyses were conducted with R version 3.5.1 in R-Studio [55]. Generalized linear mixed models (GLMMs) were run at several stages of analysis, using the packages 'nlme' [56] and 'glmmTMB' [57]. The specifics of the models used at each stage are described in the subsections below. At all stages, model assumptions were assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots [58][59]. Heteroscedasticity was avoided using the 'varIdent' variance structure transformation from the 'nlme' package when needed (or its glmmTMB-equivalent). This particular transformation allowed different residual spread for each level of the categorical variables in our model [59]. In addition, all models were run with an autoregressive correlation structure for continuous time covariates (corCAR1).

Main and interaction effects from all linear mixed models were derived using F tests with corresponding P value ($P < 0.05$). Statistical significance of random effects were computed by means of likelihood ratio tests, and reported as Chi Square values. Main and interaction effects of analyses using the package 'glmmTMB' were reported with Wald Chi Square tests, as this package does not (yet) allow extraction of F-statistics for testing. Pairwise comparisons were conducted using the package 'emmeans' [60] to follow up on main or interaction effects. To reduce the probability of a Type I error due to multiple comparisons, the α was adjusted using a Dunn-Šidák correction in all *post hoc* tests [61]. Supplementary Table S10 presents an overview of the adjusted α-value for each comparison. All *post hoc* tests were summarized as beta-estimates and their corresponding standard error, t statistic and P values. Effect sizes for *post hoc* tests were reported as Cohen's d, and obtained via the package 'emmeans'. The guidelines provided by Wahlsten [62] were used to interpret the absolute values of Cohen's d ($|d|$). This extensive review of various phenotypes suggested the following interpretation of effects for neurobehavioral mouse studies: small effect, $|d| < 0.5$; medium effect, $0.5 < |d| < 1.0$; large effect, $1.0 < |d| < 1.5$; very large effect, $|d| > 1.5$.

2.7.3 Strain differences (behavior and corticosterone)

GLMMs analyzed strain differences on each behavioral dimension using a 3 (strain) x 2 (experimenter) x 4 (trial) mixed factorial design. Strain, experimenter and trial were included as fixed predictors, as well as their two- and three-way interactions. Individual mouse (ID), slope (trial nested in ID), batch and test order were included as random effects [63]. The variables avoidance behavior, arousal, exploration and locomotion were analyzed using the package 'nlme'. The variable risk assessment was analyzed with the package 'glmmTMB' because the distribution of residuals was zero-inflated. Avoidance behavior was logarithmically transformed and locomotion rank transformed to achieve normality of the residuals. Avoidance behavior included a variance function ('varIdent') for strain (allowing different residual spread between strains) to avoid heteroscedasticity. The variables risk assessment, exploration and locomotion included the same variance function for 'trial' within 'strain' (allowing different residual spread on each trial, for each strain). Furthermore, the variable pCORT was z-transformed and pCORT levels were analyzed with a generalized least squares (gls) model using a 3 (strain) x 4 (technician) x 3 (sampling moment) mixed factorial design. Strain, sampling moment, technician and the interaction between strain and sampling moment were included as fixed predictors. Day of test was included as a covariate. Individual mouse (ID), slope, batch and test order were initially included as random factors but removed from the model because the model without random factors gave the best fit (as determined by the AIC-criterion). pCORT (nmol/L) was logarithmically transformed to achieve normality of the residuals and the variance function 'varIdent' was applied to allow different residual spread on the three samples within each strain. Detailed results of each explanatory variable, for each behavioral dimension and for pCORT, are provided in supplementary

Table S3. Significant main and/or interaction effects were followed up by *post hoc* tests. Detailed results of all *post hoc* comparisons for each dimension and for pCORT are provided in supplementary Tables S4, S5 and S6.

2.7.4 Clustering procedure

Instead of conducting the clustering procedure with the integrated z-scores of the behavioral dimensions and the z-score of pCORT, we used residual values of these z-scores for this part of the analysis. This was done as a means to control for confounding effects of strain, experimenter, batch and test order during assessment of the occurrence of sub-groups of individuals that follow a similar behavioral response across trials. Standardized Pearson residuals of the z-scores were obtained via additional LMMs using a 3 (strain) x 2 (experimenter) factorial design (behavior) or a 3 (strain) x 4 (technician) factorial design (pCORT). The factor 'trial' or 'sampling moment' was intentionally left out of the model because we wanted to maintain this information in the residuals so that we could assess the change in behavior over trials, or in pCORT over sampling moments. For each behavioral dimension and for pCORT, strain and experimenter/technician were included as fixed factors, with individual mouse (ID), batch and test order as random factors. Avoidance behavior and pCORT were logarithmically transformed to achieve normality of the residuals. Furthermore, avoidance behavior, exploration, locomotion and pCORT included a variance function for 'strain', allowing different residual spread between strains to avoid heteroscedasticity. The resulting standardized Pearson residual integrated z-scores were subsequently analyzed with a *k*-means clustering procedure using the package 'kml3d' [64] . The settings and rationale for using this particular package have been described in our previous study in detail [19]. The settings used in the present analyses are identical to those specified in [19], with the exception of the distance metric used for clustering. In the present analyses, the Fréchet distance was used because this metric is particularly sensitive for longitudinal data, whereas Euclidean distance was used in our previous study [19].

Six response trajectories were included for each individual mouse: avoidance behavior, risk assessment, arousal, exploration, locomotion and pCORT. These were clustered simultaneously to explore the occurrence of homogeneous groups of mice that followed the same response on all behavioral dimensions. Prior to analysis the gap statistic was applied to evaluate whether the trajectories were perhaps best represented by a single cluster, using the package 'cluster' [65]. This was not the case. The gap statistic compares the within-cluster sum-of-squares to a null reference distribution of the data, which is then equivalent to a single cluster [66], and as such gives an indication of whether it is appropriate to partition the data into clusters. The cluster analysis compiled 1000 iterations for each *k* clusters between 2 and 6, resulting in 5000 cluster solutions. The number of clusters was selected using the approach of Clustering Validity Indices (CVI's) [67], which was adjusted by Wahl et al. [68]. All details of this procedure are described in [19].

2.7.5 Cluster differences (behavior and corticosterone)

GLMMs analyzed cluster differences on each behavioral dimension using a 2 (cluster) x 4 (trial) mixed factorial design. A LMM analyzed cluster differences in pCORT levels using a 2 (cluster) x 3 (sampling moment) mixed factorial design. In all models cluster and trial/sampling moment were included as fixed predictors, while individual mouse and sampling time (nested in ID) were included as random factors. The variables avoidance behavior, arousal, exploration, locomotion and pCORT were analyzed using the package 'nlme' [56]. The variable risk assessment was analyzed with the package 'glmmTMB' [57] because the distribution of residuals was zero-inflated. Locomotion was rank transformed and a square root transformation was applied on pCORT to achieve normality of the residuals. The models for arousal and locomotion included a variance function ('varIdent') for cluster, allowing for differential residual spread between clusters to avoid heteroscedasticity. The models for avoidance behavior, exploration and pCORT included the same variance function for 'trial' (or 'sampling moment' for pCORT) within 'cluster' (allowing differential residuals spread on each trial/sampling moment, within each cluster). Detailed results of each explanatory variable, for each behavioral dimension and for pCORT, are provided in supplementary Table S7. Significant main and/or interaction effects were followed up by *post hoc* tests. Detailed results of all *post hoc* comparisons for each dimension are provided in supplementary Tables S8 and S9.

2.7.6 Cluster stability

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

cluster analysis was determined according to the following formula: *number of times in the same cluster/total number of bootstrapping samples*. The individual similarity indices were subsequently averaged across mice to determine the overall Jaccard similarity index for each cluster.

3. Results

3.1 Strain differences (behavior and pCORT)

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

Avoidance behavior trajectories significantly differed between strains (strain effect: $F_{(2, 111)} = 17.44$, $P < 0.0001$; trial effect: $F_{(3, 329)} = 11.04$, $P < 0.0001$; strain x trial interaction: $F_{(6, 329)} = 8.51$, $P < 0.0001$; Fig. 1, supplementary Table S3). *Post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that C decreased ($P < 0.0001$, *very large* effect size, $d = 3.312$, 95%CI [2.228, 4.396]), while 129S2 increased avoidance behavior between trial 1 and trial 4 ($P = 0.0015$, *very large* effect size, $d = -1.670$, 95%CI [-2.712, -0.628]). B6N did not display a significant change across trials (supplementary Table S4). Further *post hoc* comparisons (adjusted $\alpha = 0.016952$) showed that strains differed in onset levels of avoidance behavior, measured as mean avoidance on trial 1. C mice displayed higher onset levels of avoidance than B6N ($P = 0.0004$, *very large* effect size, $d = 2.372$, 95%CI [1.045, 3.699]) and 129S2 ($P < 0.0001$, *very large* effect size, $d = 5.595$, 95%CI [4.203, 6.987]). Furthermore, 129S2 displayed lower onset levels of avoidance behavior compared to B6N ($P < 0.0001$, *very large* effect size, $d = 3.223$, 95%CI [2.044, 4.402]), supplementary Table S5). The difference in avoidance remained significant on trial 2 between C and 129S2 (adjusted $\alpha = 0.025321$; $P < 0.0001$, *very large* effect size, $d = 2.666$, 95%CI [1.533, 3.799]) and between C and B6N ($P = 0.0057$, *large* effect size, $d = 1.439$, 95%CI [1.045, 3.699]). Overall, strain differences disappeared on trials 3 and 4 (supplementary Table S5). In addition to these strain differences, avoidance behavior scores differed between experimenters ($F_{(1, 111)} = 8.61$, $P = 0.0041$). Experimenter II (Fig. 1, right panel) scored more avoidance behavior than Experimenter I (Fig. 1, left panel, averaged over trials, $P = 0.0160$, supplementary Table S6). The size of this effect however was small ($d = -0.380$, 95%CI [-0.565, -0.196]).

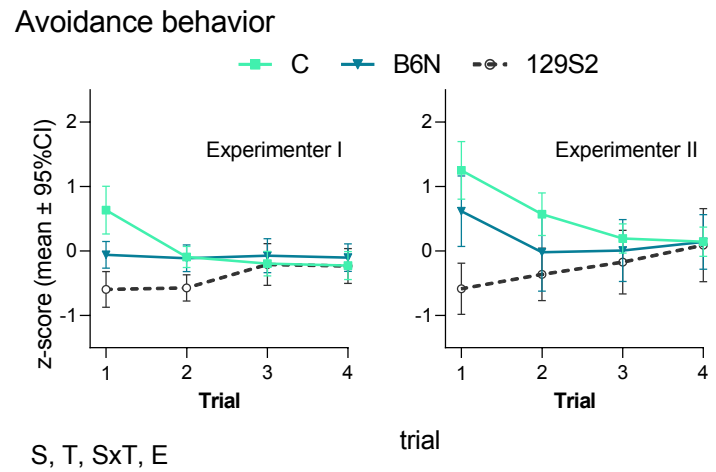


Figure 1. Display of avoidance behavior across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain;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

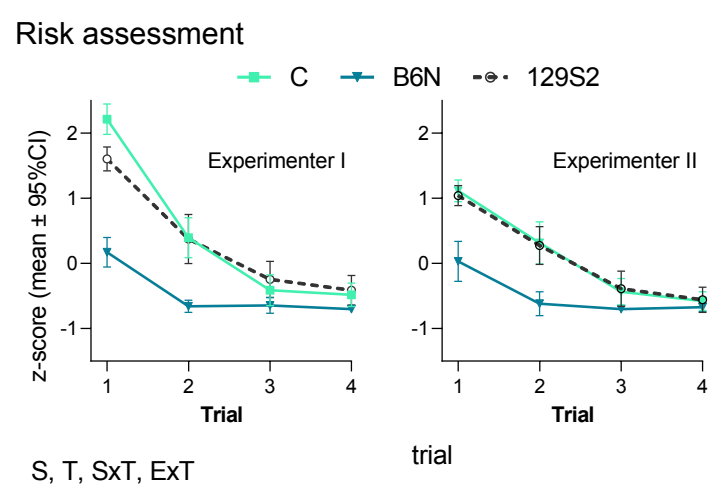


Figure 2. Display of risk assessment across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a GLMM at $P < 0.05$. S indicates a significant main effect of strain;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

Risk assessment trajectories differed significantly between strains (strain effect: $\chi^2_{(2)} = 186.53$, $P < 0.0001$; trial effect: $\chi^2_{(3)} = 678.71$, $P < 0.0001$; strain x trial interaction: $\chi^2_{(6)} = 175.72$, $P < 0.0001$), Fig. 2, supplementary Table S3. *Post hoc* comparisons (adjusted α = 0.016952) showed that all three strains displayed a significant decrease in risk assessment between the first and the last trial (C, $P < 0.0001$; B6N, $P < 0.0001$; 129S2, $P < 0.0001$, supplementary Table S4). *Post hoc* comparisons (adjusted α = 0.016952/0.025321) furthermore showed that on the first three trials, estimates of mean risk assessment were significantly lower for B6N than for C (respectively $P < 0.0001$; $P < 0.0001$; $P = 0.0014$) and lower for B6N than for 129S2 (respectively $P < 0.0001$; $P < 0.0001$; $P < 0.0001$; supplementary Table S5). C and 129S2 did not differ significantly on any of the four trials (supplementary Table S5). In addition, risk assessment trajectories differed between experimenters (experimenter x trial interaction: $\chi^2_{(3)} = 23.20$, $P = 0.0001$, supplementary Table S3). Scored risk assessment levels were significantly higher on trial 1 (adjusted α = 0.025321; $P < 0.0001$, supplementary Table S4) for Experimenter I (left panel) than for Experimenter II (right panel). Experimenters did not differ in scored risk assessment behavior on the remaining trials (supplementary Table S6). Overall, the effect size for this experimenter effect was negligible ($d = 0.128$, 95%CI [-0.055, 0.311]).

Arousal trajectories differed between strains (strain effect: $F_{(2, 111)} = 3.49$, $P = 0.0339$; trial effect: $F_{(3, 329)} = 20.66$, $P < 0.0001$; strain x trial interaction: $F_{(6, 329)} = 3.57$, $P = 0.0019$; Fig. 3, supplementary Table S3). *Post hoc* comparisons (adjusted α = 0.016952) showed that all three strains significantly increased arousal between the first and the last trial (C, $P = 0.0015$, *moderate* effect size, $d = -0.784$, 95%CI [-1.272, -0.298]; B6N, $P = 0.0006$, *moderate* effect size, $d = -0.888$, 95%CI [-1.397, -0.379]; 129S2, $P < 0.0001$, *very large* effect size, $d = -1.654$, 95%CI [-2.169, -1.140]; supplementary Table S4). Arousal levels however were highly similar between strains, as post hoc comparisons only showed a significant difference in arousal on trial 2, with higher levels of arousal in B6N compared to 129S2 (adjusted α = 0.025321; $P = 0.0002$, moderate effect size, $d = 0.931$, 95%CI [0.436, 1.426]). On the remaining trials, arousal did not differ significantly between strains (supplementary Table S5). Finally, scored levels of arousal differed between experimenters (experimenter effect: $F_{(1, 111)} = 9.15$, $P = 0.0031$, Fig. 3, supplementary Table S3). Arousal scored by Experimenter I (Fig. 3, left panel) was statistically higher than that of Experimenter II (Fig. 3, right panel, $P = 0.0007$, *small* effect size, $d = 0.409$, 95%CI [0.224, 0.594], supplementary Table S6).

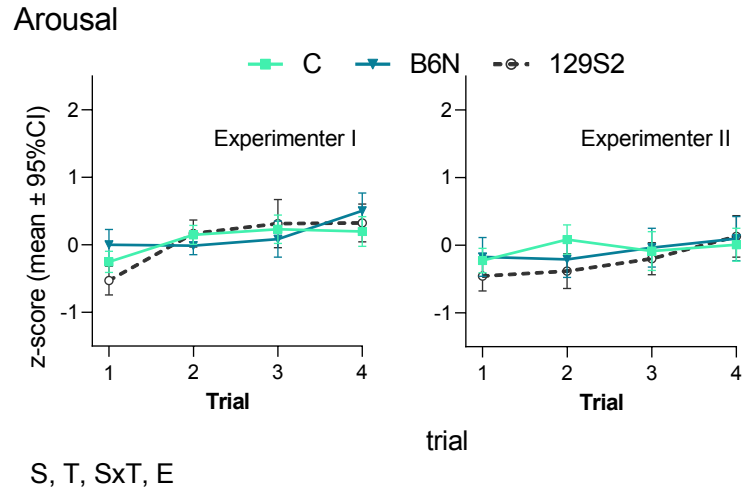


Figure 3. Display of arousal across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates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

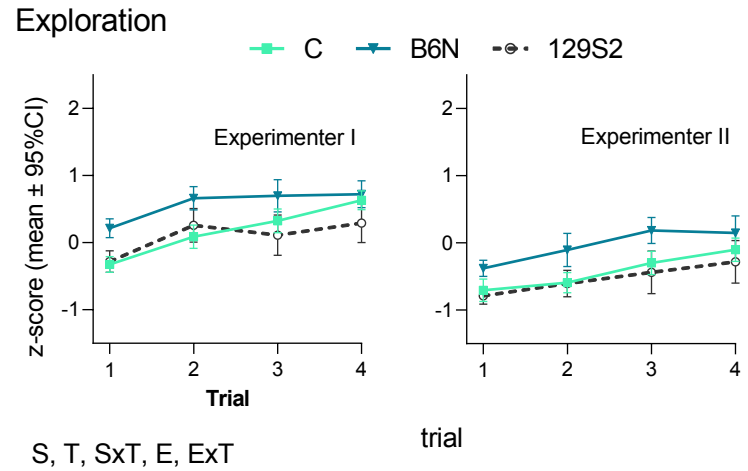


Figure 4. Display of exploration across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates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

Exploration trajectories differed significantly between strains (strain effect: $F_{(2, 111)} = 38.15$, $P < 0.0001$; trial effect: $F_{(3, 329)} = 20.66$, $P < 0.0001$; strain x trial interaction: $F_{(6, 329)} = 3.73$, $P = 0.0113$, Fig. 4 and supplementary Table S3). *Post hoc* comparisons (adjusted α = 0.016952) indicated that all three strains significantly increased exploration between the first and the last trial (C, $P < 0.0001$, *very large* effect size, $d = -2.591$, 95%CI [-3.070, -2.113]; B6N, $P < 0.0001$, *very large* effect size, $d = -1.701$, 95%CI [-2.208, -1.194]; 129S2, $P < 0.0001$, *very large* effect size, $d = -1.787$, 95%CI [-2.432, -1.142], supplementary Table S4). *Post hoc* comparisons (adjusted α = 0.016952/0.025321) further showed that B6N displayed higher levels of exploration than 129S2 on all four trials, and higher levels of exploration than C mice on the first three trials (supplementary Table S5). Mean exploration did not differ between C and 129S2 mice on any of the trials (supplementary Table S5). In addition, exploration trajectories also differed between experimenters (experimenter effect: $F_{(1, 111)} = 114.95$, $P < 0.0001$; experimenter x trial interaction: $F_{(3, 329)} = 5.08$, $P = 0.0019$, Fig. 4, supplementary Table S3). *Post hoc* comparisons (adjusted α = 0.025321/0.05) comparing exploration scores between experimenters on each trial showed that observed exploration was significantly higher for Experimenter I than for Experimenter II on all four trials (Fig. 4, supplementary Table S6). These differences were accompanied by very large effect sizes on all trials (supplementary Table S6).

The model for locomotion (rank transformed) showed that the trajectories for locomotion differed significantly between strains (strain effect: $F_{(2, 111)} = 183.02$, $P < 0.0001$; trial effect: $F_{(3, 329)} = 18.91$, $P < 0.0001$; strain x trial interaction: $F_{(6. 329)} = 20.94$, $P < 0.0001$; Fig. 5, supplementary Table S3). *Post hoc* comparisons (adjusted α = 0.016952) showed that B6N significantly decreased, while C significantly increased locomotion between trial 1 and trial 4 (B6N, $P < 0.0001$, *very large* effect size, $d = 1.715$, 95%CI [1.375, 2.054]; C, $P < 0.0001$, *moderate* effect size, $d = -0.971$, 95%CI [-1.421, -0.520], supplementary Table S4). 129S2 did not display a significant change in locomotion between these trials (supplementary Table S4). Furthermore, locomotion was significantly higher for B6N than for 129S2 on all four trials, and higher than C on the first three trials (supplementary Table S5). In addition, locomotion was also significantly higher for C mice than for 129S2 on all four trials (supplementary Table S5). Locomotion scores per strain also differed between experimenter (strain x experimenter interaction: $F_{(2, 111)} = 3.50$, $P = 0.0336$, supplementary Table S3). *Post hoc* comparisons indicated that overall locomotion scores for C mice were higher for Experimenter I than for Experimenter II ($P = 0.0232$, moderate

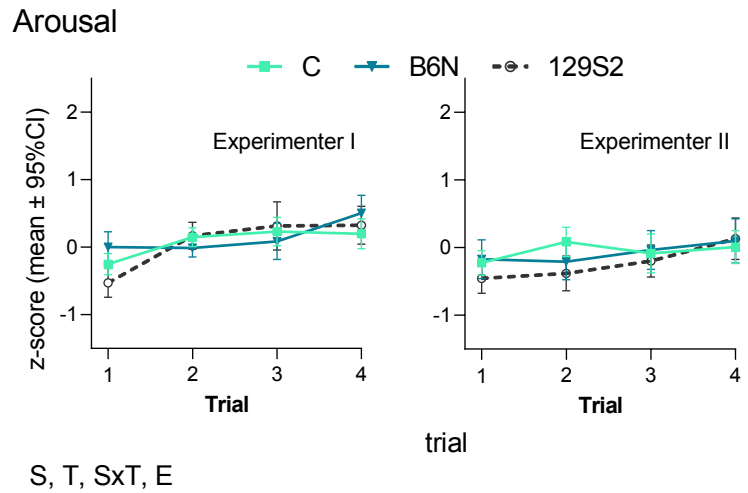


Figure 5. Display of locomotion across trials in strains C, B6N and 129S2, as scored by each experimenter. Results are expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S indicates a significant main effect of strain; T indicates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

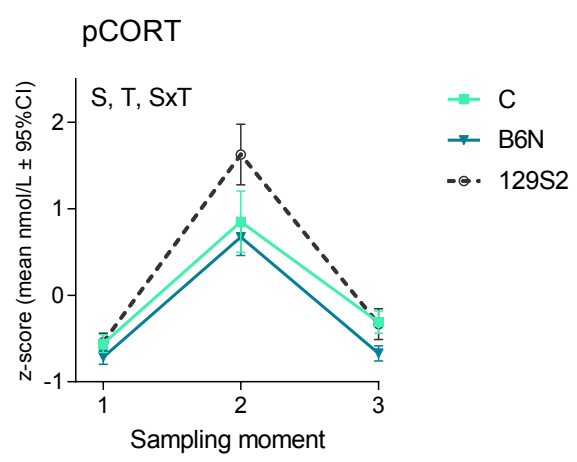


Figure 6. Blood plasma corticosterone (pCORT) levels in strains C, B6N and 129S2 one week prior to behavioral test (sampling moment 1), directly after behavioral test (sampling moment 2) and one week after behavioral test (sampling moment 3). Results are expressed as z-transformed nmol/L and presented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

effect size, d = 0.588, 95%CI [0.076, 1.100], Fig. 5, supplementary Table S6). Experimenters scores of locomotion were nog significantly different for 129S2 and B6N (supplementary Table S6).

The trajectories of pCORT differed significantly between strains (strain effect: $F_{(2, 321)}$ = 21.81, P < 0.0001; sampling moment effect: $F_{(2, 321)}$ = 263.98, P < 0.0001; strain x sampling moment interaction: $F_{(4, 321)}$ = 4.67, P = 0.0011; Fig. 6, supplementary Table S3). *Post hoc* comparisons (adjusted α = 0.012741) revealed that pCORT levels were higher on sampling moment 2 than baseline (sampling moment 1) for all three strains (C, P = 0.0009, *large* effect size, d = -1.389, 95%CI [-2.194, -0.585]; B6N, P = 0.0002, *very large* effect size, d = -1.676, 95%CI [-2.541, -0.812]; 129S2, P = 0.0002, *very large* effect size, d = -2.460, 95%CI [-3.701, -1.219]; supplementary Table S4). Furthermore, *post hoc* comparisons (adjusted α = 0.012741) showed that strains did not significantly differ in baseline pCORT levels (supplementary Table S5). Directly after behavioral test however (sampling moment 2), pCORT was significantly higher for 129S2 mice than for B6N (P = 0.00281, *moderate* effect size, d = -0.979, 95%CI [-1.694, -0.264]) and there was a suggestive effect of higher pCORT in 129S2 compared to C (P = 0.0104, *large* effect size, d = -1.102, 95%CI [-1.937, -0.267]; supplementary Table S5). A week after behavioral test (sampling moment 3), pCORT had decreased significantly in all three strains compared to sampling moment 2 (C, P = 0.0028, *large* effect size, d = 1.069, 95%CI [0.375, 1.762]; B6N, P = 0.0003, *very large* effect size, d = 1.616, 95%CI [0.772, 2.460]; 129S2, P = 0.0002, *very large* effect size, d = 2.273, 95%CI [1.110, 3.435], supplementary Table S4). On this sampling moment however, pCORT levels were significantly lower for B6N than for C (P = 0.0030, small effect size, d = 0.424, 95%CI [0.151, 0.698]; supplementary Table S5). Thus, in all three strains pCORT levels were significantly higher directly after behavioral test, compared to a week before test, with 129S2 presenting the highest pCORT levels directly after test compared to the other two strains. One week after behavioral testing, pCORT levels had decreased significantly in all strains, with C being the only strain with significantly higher pCORT levels on sampling moment 3, compared to baseline on sampling moment 1 (adjusted α = 0.01695; P = 0.0077, small effect size, d = -0.320, 95%CI [-0.555, -0.086], supplementary Table S4).

3.2. Cluster analysis

Standardized Pearson residuals of the integrated z-scores were used for the clustering of the individual trajectories (see Section 2.7.4). Data from all strains was pooled in order to assess individual variation in habituation responses within and across strains. All five behavioral dimensions and pCORT were taken into account simultaneously. As such, six response trajectories were included for each individual mouse: avoidance behavior, risk assessment, exploration, locomotion, arousal and pCORT. The optimal partitioning of the data yielded two clusters. Table 1 presents cluster size and distribution of strains across clusters. The mice were more or less evenly distributed across clusters, with 53.8% of mice (n = 63) grouping together in cluster A while the remaining 46.2% (n = 54) fell in cluster B. Cluster A and cluster B both consisted of individuals from all three strains, but the distribution of strains differed between clusters. The majority of C (82.5%, n = 33 out of 40) fell in cluster A, while the majority of 129S2 (76.3%, n = 29 out of 38) grouped together in cluster B. B6N mice were divided over clusters A (53.8%, n = 21) and B (46.2%, n = 18). Because of the marked experimenter effects in the strain analyses, the distribution of mice within clusters are presented for each experimenter separately.

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

Cluster size (n) and proportion of total n per cluster		
	Cluster A	**Cluster B**
n total = 117	*n* = 63 *(53.8%)*	*n* = 54 *(46.2%)*

Distribution of strains *within* clusters and per experimenter

	Cluster A			**Cluster B**		
	Exp. I	Exp. II	Total	Exp. I	Exp. II	Total
Strain	*n* (%)	*n* (%)	n (%)	*n* (%)	*n* (%)	*n* (%)
C	16 (48.5)	17 (51.5)	33 (52.4)	4 (57.1)	3 (42.9)	7 (13.0)
B6N	12 (57.1)	9 (42.9)	21 (33.3)	12 (66.7)	6 (33.3)	18 (33.3)
129S2	3 (33.3)	6 (67.7)	9 (14.3)	16 (55.2)	13 (44.8)	29 (53.7)

3.3. Cluster differences (behavior and pCORT)

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

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

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

Furthermore, clusters differed significantly with respect to activity related dimensions. Locomotion (rank transformed) was significantly higher in cluster B compared to cluster A regardless of trial (cluster effect: $F_{(3,341)}$ = 11.35, P = 0.0010; Fig. 7, supplementary Table S7). Exploration trajectories differed between clusters (trial effect: $F_{(3, 341)}$= 59.73, P < 0.0001; interaction cluster x trial: $F_{(3,341)}$ = 8.59, P < 0.0001; Fig. 7, supplementary Table S7). *Post hoc* comparisons (adjusted α = 0.02532) however indicated that both clusters increased exploration between trial 1 and trial 4 (cluster A, P < 0.0001, *very large* effect size, d = -2.949, 95%CI [-2.905, -2.083]; cluster B, P = 0.0001, *medium* effect size, d = -0.935, 95%CI [-1.348, -0.539]; supplementary Table S8). The significantly higher exploration levels in cluster A compared to B on trials 3 and 4 however suggest the increase in exploration was more pronounced in cluster A (trial 3, P = 0.0458; *medium* effect size, d = 0.647, 95%CI [0.007, 1.287; trial 4, P = 0.0023, *large* effect size, d = 1.065, 95%CI [0.374, 1.757]; Fig. 7, supplementary Table S9).

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

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

3.4. Relative weight of dimensions on cluster partitioning

The obtained clusters were based on simultaneous clustering of all five behavioral dimensions and pCORT. However, cluster differences were more pronounced on some variables than on others, with significant differences between clusters in avoidance behavior, arousal, exploration and locomotion, but not in risk assessment and pCORT levels. In order to assess the relative impact of each variables on this partitioning, we conducted additional cluster analyses, each time leaving one of these six dimensions out. Pearson Chi square tests showed that cluster size in any of these analyses did not significantly differ from cluster size in the original cluster analysis (Table 2). The Jaccard similarity index subsequently indicated how many individual mice were retained in the same cluster as in the original cluster analysis, after excluding a certain behavioral dimension. As can be seen in Table 2, only 53% of the mice retained their cluster after omitting avoidance behavior, while omitting any of the other

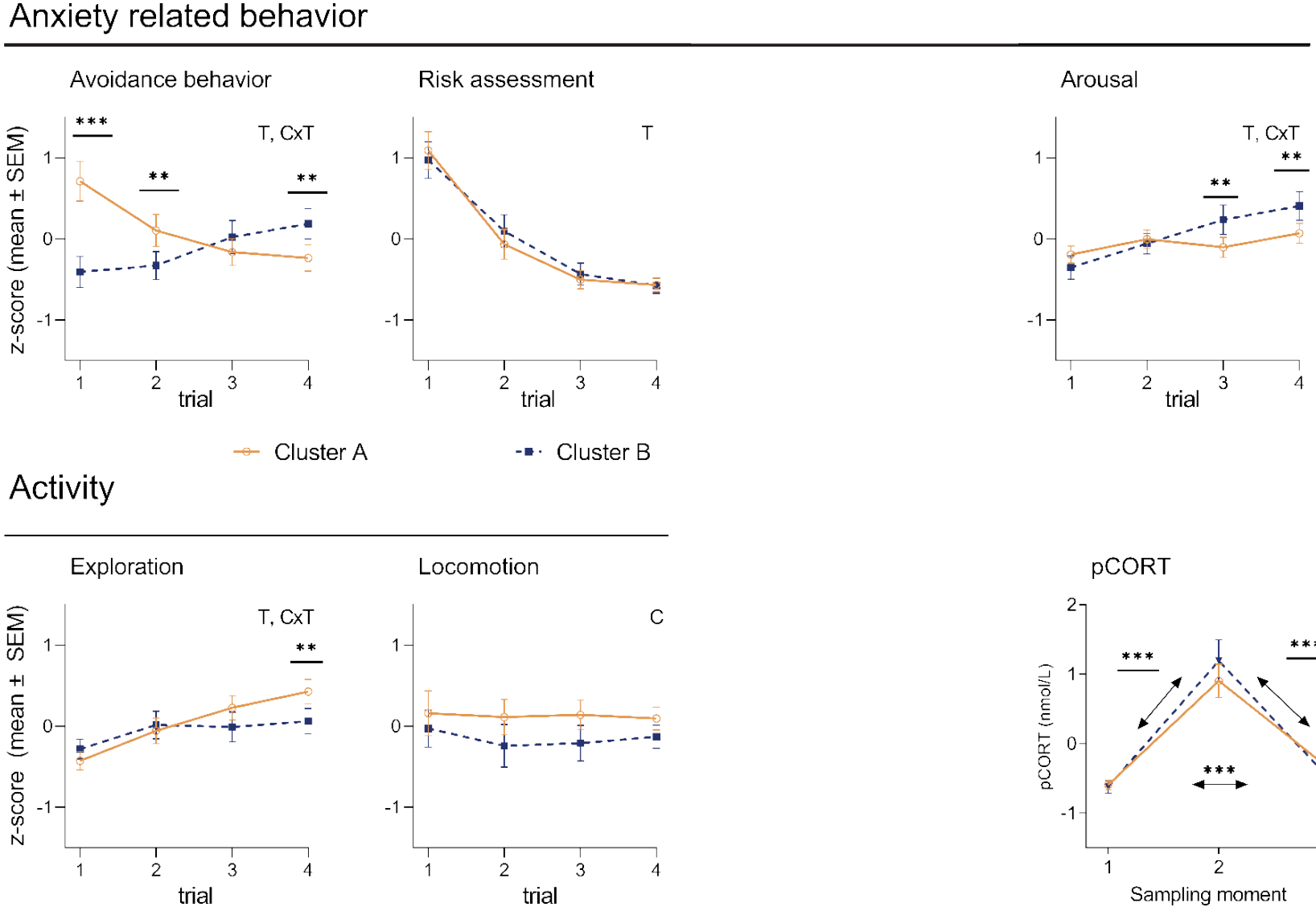


Figure 7. Differences between clusters on each behavioral dimension, and corticosterone levels. Behavior expressed as integrated behavioral z-scores for behavioral dimensions, and the z-score nmol/L for pCORT. Results are presented as means with 95% CI. Effects were significant in LMMs at $P < 0.05$.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

dimensions hardly affected cluster membership for individual mice (Jaccard indices > 0.90, Table 2). Thus, avoidance behavior appeared most dominant in partitioning of the clusters.

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

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

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

3.5. Cluster stability

Fig. 8 depicts the mean trajectory of all bootstrap samples (black dashed line) against the trajectory belonging to the original cluster (cluster A, orange; cluster B, blue), as well as the 200 trajectories of the bootstrap samples (grey), for each cluster, on each dimension. For cluster A, the average Jaccard similarity index was 0.64, meaning that on average, an individual mouse belonged to cluster A in 64% of the bootstrap samples. The average Jaccard similarity index for cluster B was 0.53.

3.6. Cluster differences within strain (behavior and pCORT)

Finally, each cluster consisted of individuals of all three strains. These strains showed quite distinct behavioral profiles, of differential adaptive value. Fig. 9 therefore provides a visual representation of how the identified clusters were expressed within each strain separately. LMMs analyzed cluster differences for each dimension within each strain using a 2 (cluster) x 4 (trial) mixed factorial design. Cluster, trial and their interaction were included as fixed predictors, and individual mouse as random factor. To avoid repetition of statistical results, the specification of cluster differences within each strain in the section below is purely descriptive. Detailed results of all analyses are presented in supplementary tables S11 and S12. For a quick reference, significant main and interaction effects, and any significant *post hoc* tests comparing cluster differences per trial are depicted in Fig. 9.

3.6.1. C

On average, C mice successfully habituated to the test. Within this strain however, subgroups of mice differed significantly in avoidance behavior and arousal (Fig. 9, Supplementary Table S11). The majority of C (n = 33) grouped together in cluster A (Table 2). The behavioral profile of this cluster was indeed highly similar to the average C response: initial high levels of avoidance behavior and risk assessment that significantly decreased, while arousal, exploration and locomotion significantly increased across trials (Fig. 9, top row; Supplementary Table S12). A small subgroup of C (cluster B, n = 7) however, displayed significantly lower levels of avoidance behavior than cluster A on the first trial, and higher levels of avoidance behavior on trial 4 (Fig. 9, top row; Supplementary Table S12). In addition, avoidance behavior remained stable across trials (Supplementary Table S12). This subgroup also displayed a more rapid increase in arousal compared to their counterparts in cluster A, with significantly higher levels of arousal on the last two trials (Fig. 9, top row, Supplementary Table S12).

3.6.2. B6N

On average, B6N were characterized by high activity, and low anxiety levels. This strain was distributed almost equally across clusters, with n = 21 individuals in cluster A, and n = 18 individuals in cluster B (Table 2). Within B6N, these clusters differed significantly on avoidance behavior, arousal, exploration and locomotion (Fig. 9, middle row; Supplementary Table S11). On average, avoidance behavior remained stable across trials in B6N. The identified clusters however displayed contrasting patterns of avoidance behavior across trials: a significant decrease in cluster A, and a significant increase in cluster B between the first and the last trial (Fig. 9, middle row; Supplementary Table S12), with significantly higher levels of avoidance behavior on trial 1 for B6N mice in cluster A (Supplementary Table S12). Arousal increased in both clusters for B6N, but arousal was higher on the last two trials in cluster B compared to cluster A (Fig. 9, middle row; Supplementary Table S12). Exploration did not differ between clusters on any of the trials, and overall locomotion was higher in B6N mice in cluster A than in cluster B (Supplementary Table S12).

3.6.3. 129S2

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

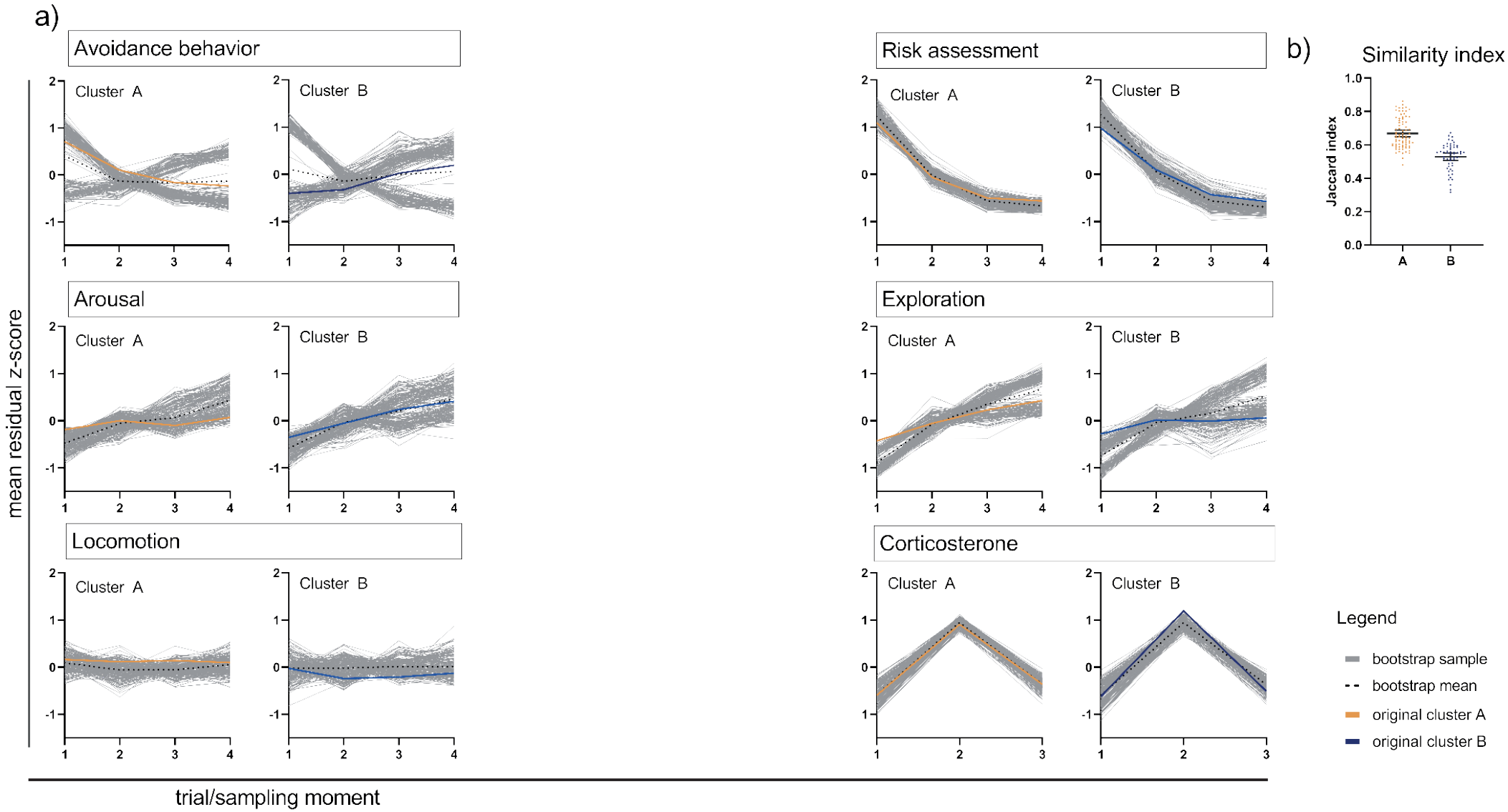


Figure 8. Results of the bootstrapping procedure for each cluster, on each behavioral dimension. Results are presented as mean residual z-scores, and depict the trajectory of the original cluster (cluster A, orange; cluster B, blue) in relation to the average of all 200 bootstrap samples (black dashed line), against all 200 trajectories of the bootstrapping procedure (grey). **(b)** Distribution of individual Jaccard Indices in clusters A and B (cluster A, orange dots; cluster B, blue dots). Average Jaccard Index per cluster indicated by mean with 95% CI (black).

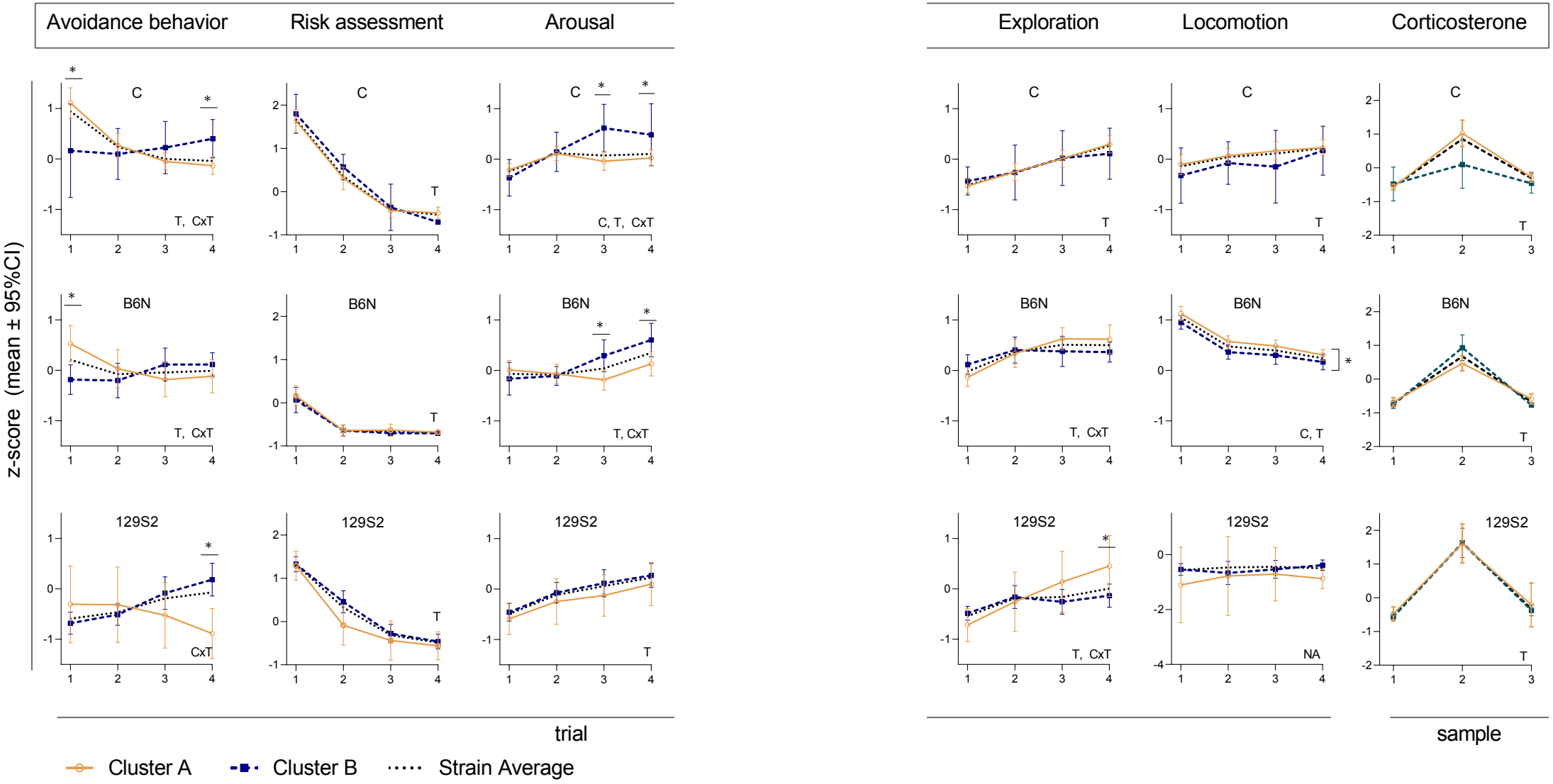


Figure 9. Differences between clusters within strains (C, top row; B6N, middle row; 129S2, bottom row) on each behavioral dimension, and corticosterone levels. Behavior expressed as integrated behavioral z-scores for behavioral dimensions, and the z-score nmol/L for pCORT. Results are presented as means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

129S2 grouped together in cluster B (n = 28, 63.2%), while the remaining individuals fell in cluster A (n = 9, 36.8%). Within this strain, the two clusters differed significantly on avoidance behavior and exploration (Fig. 9, bottom row; Supplementary Table S11). Avoidance behavior in cluster B increased significantly, while this behavior decreased in 129S2 mice in cluster A, with significantly lower avoidance behavior on trial 4 in cluster A than in cluster B (Fig. 9, bottom row; Supplementary Table S12). Exploration trajectories differed between clusters, with significantly higher exploration on trial 4 for 129S2 mice in cluster A (Fig. 9, bottom row; Supplementary Table S12).

4. Discussion

We showed that behavioral behavioral responses to novelty, measured as habituation or sensitization of anxiety-related behavior in the mHB, differ between individuals in three commonly used mouse inbred strains. The behavioral profiles of the identified clusters were largely similar to the habituation and sensitization profile previously identified by van der Goot et al. [19]. As noted in the introduction, we asked whether these previously identified clusters were perhaps (partly) result of the variation that was inherent to re-analyzing a dataset consisting of multiple experiments. This aggregated dataset varied naturally in factors that are known to affect variability between experiments, such as test location, time of year, experimenter, etcetera [20] [21]. The consistency in behavioral profiles between [19] and the present experiment however suggests that this was not the case. As such this study empirically confirms that adaptive capacities may be subject to inter-individual variability in 129 mice. Also in line with van der Goot et al. [19], the identified response types were multidimensional, with individual mice grouping together across both anxiety and activity related behavioral dimensions. This confirms the notion that anxiety is a complex behavioral construct that is expressed by multiple behavioral dimensions [18][27][71].

At the same time, the differential behavioral profiles were not substantiated with differential endocrine profiles. Glucocorticoid responses have been associated with anxiety-related behavior in mice [28][70], and have been found to vary greatly between individuals in rodents [29][30][72][73]. Previous findings showed that differential behavioral response types may be correlated to differences in endocrine response in inbred strains [32][33][34]. Jakovcevski et al. [32] for example identified sub-populations of B6 exhibiting high and

low trait anxiety (measured as the latency to freely enter a novel environment from their home cage) and showed that trait anxiety was positively correlated with pCORT concentrations after exposure to a stressor [32]. This correlation however was highly dependent on the type of stimulus used. The association was only found after individuals were exposed to a severe stressor (exposure to a rat), and not when mice were exposed to a mild stressor, a novel environment [32]. Exposure to novelty, a major characteristic of the paradigm employed in our experiment, is indeed typically regarded as a mild stressor [74] and this may have affected our results. On a behavioral level however, clustering individual response trajectories yielded two multidimensional subtypes of response (Fig. 7). In line with [19], avoidance exerted the most 'weight' on the partitioning of the clusters, suggesting that this behavioral dimension is an important distinguishing factor of these multidimensional response types (Table 2).

New to the findings by van der Goot et al. [19] was that these behavioral subtypes were not only displayed by 129 mice, but also present in C and a newly included strain, B6N. These three strains are known for their differential innate emotionality [17], and this should be taken into account when interpreting the differential response types. The highly neophobic [18] but adaptive phenotype that is characteristic for C for example, [11][14][15][16] was also observed in the present study. On average C mice displayed initial high levels of anxiety-related behavior that decreased rapidly, while initial low levels of activity increased with trials. The individual based analyses demonstrated that the majority of C indeed grouped together in cluster A, which mirrored the adaptive profile of C mice (*n* = 33, 82.5%, Table 1). A small subgroup however deviated from this response with low initial levels of avoidance behavior and a more pronounced increase in arousal, while overall levels of locomotion were lower than their counterparts in cluster B (*n* = 7, 17.5%, Table 1, Fig. 9). In general, C mice are often characterized as phenotypically robust, showing relatively little within strain variation compared to other inbred strains [75]. At the same time, subtypes in emotional reactivity and sensitivity to stress have been identified previously in these animals [76]. The seemingly less adaptive profile could suggest that such inter-individual variability may also pertain to the adaptive anxiety phenotype that is characteristic for this strain. At the same time, it should be kept in mind that the small number of C mice that deviate from their average strain response may equally represent individuals that are less responsive to the test, and further assessment of these subtypes, for example by means of pharmacological validation, could provide more insight on this.

Next, the non-adaptive phenotype that is characteristic for various sub-strains of 129 mice [11][13][14][15][16] was confirmed in the present study. On average avoidance behavior increased in 129S2, while low levels of locomotion remained stable across trials and exploration increased. In addition, average pCORT levels were significantly higher in 129S2 mice directly after behavioral testing compared to C- and B6N mice. This finding in itself was interesting because it has been speculated before, that the lack of habituation in 129S2 mice (i.e. a lack of exploration of the unprotected area in the mHB) might be due to persistent low levels of locomotor activity and not related to anxiety-related characteristics [13]. Sub-strains of 129S2, including 129S2/SvPasCrl tested here, are indeed known for their reduced activity levels [26][77][78]. That pCORT levels were highest in 129S2 mice however, suggests that exposure to the mHB was indeed perceived as particularly stressful by this strain, and could serve as an indication that the anxiety phenotype of 129S2 can be classified as non-adaptive. At the same time, there are many additional factors that affect corticosterone levels at a given time point (i.e. circadian rhythm [79], intestinal microbiome [80], nutritional status [81]), so further research is necessary to determine whether the elevated corticosterone levels in 129S2 mice were indeed associated with increased anxiety or whether other factors were at play. In line with [19], the majority of 129S2 mice grouped together in the cluster that mirrored the average behavioral non-adaptive 129S2 anxiety phenotype (Cluster B, n = 29, 76.3%, Table 1). Also in line with this study, in a small subgroup of 129S2 mice (cluster A, n = 9, 23.7%, Table 1) anxiety-related behavior decreased across trials, suggesting a seemingly more adaptive phenotype (Fig. 7), with lower levels of avoidance behavior on the last trial, lower risk assessment, risk assessment and arousal (all three not significant) and a more pronounced increase in exploration, than their counterparts in cluster B (Fig. 9). This could suggest that not all 129S2 individuals are equally susceptible to the aversive nature of the mHB.

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

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

The fact that on average B6N mice displayed a low anxious, highly active profile, and the fact that further zooming in on these individual responses showed that within B6N, avoidance behavior only differed significantly on the first trial between clusters (Fig. 9), raises the question, whether the identified subtypes in B6N in the mHB should indeed be interpreted as differential anxiety phenotypes. B6 is often used as a comparator strain when testing for anxiety [83]. Despite its reputation as a low anxious strain however, inter-individual variability in anxiety related behavior has repeatedly been demonstrated [32][33][76][84][85]. In fact, the majority of studies that identified subtypes of emotional reactivity within inbred strains have been conducted with B6-mice. These distinct anxiety-profiles in B6 mice have been found consistent across time and context [84][85], and have even been linked to copy number variation in small nucleolar RNA clusters, suggesting a genetically specified modulator of these differential profiles [85]. At the same time, one could interpret the identified subtypes as indicative of differential activity levels, rather than indicative of differential anxiety profiles. The two clusters indeed differed in exploration and locomotion and inter-individual variability in activity related behavior has been established previously in B6 mice [86]. The existing literature does not provide a definitive explanation in either direction. Further zooming in on these subtypes is necessary to evaluate to what extent these profiles of B6 mice are related to differential activity levels, or whether they also reflect differential anxiety responses.

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

activity [13][87] and thus independent from anxiety. The lower activity levels of the small sub-population of C mice could simply represent individuals that were less responsive to the aversive nature of the mHB. Conversely, the higher activity levels in the small sub-group of 129S2 individuals compared to the majority of 129S2 mice could have affected the observed decrease in avoidance behavior. Locomotion has repeatedly been dissociated from avoidance behavior by factor analyses on the behavioral variables observed in the mHB [27][51]. Labots et al. [87] however also observed that sub-strain differences in avoidance behavior disappeared after controlling for horizontal locomotor activity (i.e. the type of locomotor activity that was recorded in the present study), suggesting that that locomotion may in fact exert a confounding effect on avoidance behavior in the mHB. In the present study, differences in avoidance behavior between clusters remained intact when including locomotor activity as a covariate in the model (trial effect: $P < 0.0001$; interaction strain x trial: $P < 0.0001$). The current results unfortunately do not allow for a definitive dissociation between anxiety-related behavior and a potentially confounding effect of locomotion. Further research should therefore first be aimed at ruling out this potentially confounding effect, for example by assessing whether the observed subtypes respond differently to pharmacological treatment, or by combining the currently employed assay with a behavioral test that is less dependent on locomotor activity, such as the physiological anxiety paradigm stress-induced hyperthermia [89]. Further pharmacological validation may also provide more insight in the question whether the identified subtypes indeed reflect two qualitatively differential anxiety responses, or whether these sub-populations represent individuals that were simply were less responsive to the test. As noted above, given the strain specific differences in emotionality, such further assessment is ideally determined within each strain separately.

When further evaluating the potential differential anxiety profiles in these strains, it should furthermore be assessed whether the identified variation is consistent across time and contexts [6][9]. As described earlier, the advantages of individual-based characterization are considered twofold: It may enable the selection of susceptible or (un)responsive individuals from a cohort – provided that sufficient time is allowed before retesting in the same or other paradigms [89] - and could thereby make animal models more representative. Second, identifying subtypes of response could provide a starting point for the exploration of the biological mechanisms underlying these subtypes [6]. These presumed benefits are however highly dependent on the temporal consistency of the behavior of interest [6][9][85]. Consistency of anxiety-related

and activity behaviors across time [84][85][86] and contexts [84] has so far only been demonstrated in B6 mice (though not in the mHB).The consistency of the behavioral profiles between the present study and van der Goot et al. [19] give a first indication that subtypes exist in at least 129 mice but this requires further study.

Thus, although further research is necessary to determine whether the identified subtypes reflect differential anxiety profiles, or represent differential responsivity to the test, the present results show that mice of various inbred strains may differ in their behavioral anxiety response. This finding in itself is also relevant from another perspective, namely that defining animals on an individual level and incorporating this information in the analysis of results may contribute to the quality of animal experiments [19]. Lonsdorf and Merz [5] for example argued that subpopulations displaying contrasting response patterns may obscure the detection of significant differences at group level (i.e. a type II error). Moreover, incorporating inter-individual variability in the design and analysis may enable researchers to better adhere to one of the fundamental principles of good experimental design of animal experiments: that all variables should be controlled, except that due to treatment [92]. Inter-individual variability – with its complex origin and elusive nature – has been proven a major factor undermining this principle [8]. This complexity makes it challenging to completely control for this type of variability through increased standardization and therefore an increasing body of research has advocated the incorporation of this variability in the design and statistical analysis of animal experiments as an alternative way forward [93][94][95][96]. The presently applied approach as such may contribute to existing approaches advocating such incorporation.

With the benefits of this study stated, the findings also presented a number of limitations, which are discussed below. For one, the cluster stability of the presently identified clusters was low compared to the stability of the clusters identified by van der Goot et al. [19], despite the fact that the behavioral profiles largely overlapped between the two studies. This was unexpected, especially since we found highly stable clusters in a second (to be published) dataset obtained from testing the same strains in the same behavioral test. The instability of the clusters at first glance contradicts the suggestion that the identified profiles reflect inter-individual variability in the observed strains. An alternative explanation may however be found in the intricate properties of k-means cluster analysis. In unsupervised cluster analyses, the stability of

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

Second, the present study also shows how experimenter effects can affect experimental results. The experimenter has been widely acknowledged as an uncontrollable factor in animal experimentation [20][96][98][99][100]. Factors such as sex [101], familiarity with the experimenter [102] and experimenter experience [96] may all affect behavioral traits. Inter-observer variability constitutes another form of experimenter-induced variation [96][103]. In this study, the kappa statistic indicated moderate to good inter-observer reliability with on average a high percentage agreement (Section 2.4). Inter-observer reliability in itself however can again be affected by numerous factors such as experience, training, the rapidity of behavior, energy level of the observer and so on [96][104]. Automated tracking has been advocated as a means to overcome the incontrollable nature of this 'human element' and as such to increase the standardization of an experiment [103]. To our knowledge, fully automated scoring has unfortunately not yet been validated in the modified Hole Board, and doing so was beyond the scope of the present study. Furthermore, each experimenter was allocated to one of the two animal rooms making it difficult to dissociate between experimenter-related and room effects. As described in section 2.2, the humidity and temperature were comparable between the two animal rooms, but other factors could have played a role as well (i.e. barometric pressure, noise) [105]. A means to dissociate between these experimenter-related effects and room-effects in future research would be to have both experimenters observe data in both animal rooms. Although the factor experimenter was controlled for in the residuals that were used for cluster analysis, the current study does not permit a definitive exclusion of the

possibility that the found experimenter effects affected (some of the) variability that was found in the data. If anything, the experimenter effects in the present study emphasize the importance of accounting for this factor in experimental design, analysis and report of the results [97].

Third, the fact that the behavioral test was conducted in the housing room may have affected variability between individuals that were tested later in the experiment in comparison to animals that were assessed earlier on in the experiment. Research has demonstrated that ultra-sonic vocalizations may affect corticosterone responses and behavior in mice [106]. We have attempted to avoid a bias in our data of this potential confounding influence by randomizing test order across strains, within test day and batch. Furthermore, the random factor 'test order' did not contribute significantly to the variance in the models that assessed between strain differences, and test order was controlled for when obtaining the residuals that were used for clustering the data.

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

5. Conclusion

This study empirically demonstrates, that inter-individual variability in habituation and sensitization of anxiety responses exists within males of three commonly used mouse inbred strains. The currently identified profiles are in line with previous findings, and as such suggest that they may be representative of subtypes of behavioral response in the observed strains. The three strains differ in innate emotionality. Whether the identified sub-groups represent differential

adaptive capacities regarding anxiety responses, or whether they represent individuals that are less responsive to the test may therefore differ between strains and requires further study. Also, further study is required to map inter-individual variability in female mice of these strains. The profiles identified in this study however provide a useful starting point for such further assessment.

Acknowledgements

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

References

1. Cohen, H., Geva, A. B., Matar, M. A., Zohar, J., Kaplan, Z., 2008. Post-traumatic stress behavioural responses in inbred mouse strains: can genetic predisposition explain phenotypic variability? Int. J. Neuropsychoph. 11, 331-349, https://dx.doi.org/10.1017/S1461145707007912.
2. 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, http://doi.org/10.1016/j.yfrne.2010.04.001.
3. 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. http://doi.org/ 10.3389/fphar.2013.00137.
4. 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.
5. Lonsdorf, T. B., Merz, C. J., 2017. More than juist noise: Inter-individual differences in fear acquisition, extinction and fear in humans – Biological, experiential, temperamental factors, and methodological pitfalls. Neurosci. Biobehav. Rev. 80, 703-728, https://doi.org/10.1016/j.neurobiorev.2017.07.007.
6. 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.
7. Lathe, R. (2004). The individuality of mice. Genes Brain Behav. 3, 317–327, https://doi.org/10.1111/j.1601-183X.2004.00083.x.
8. Voelkl, B., Altman, N.S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N.A., Kas, M.J., Schielzeth, H., Van de Casteele, T., Würbel, H. 2020. Reproducibility of animal research in the light of biological variation. Nat. Rev. Neurosci. 2020; 21, 384-393, https://doi.org/10.1038/s41583-020-0313-3.
9. 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
10. 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.
11. 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, http://doi.org/10.1111/j.1601-183X.2009.00527.x

12. 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.
13. 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, http://doi.org/10.1371/journal.pone.0042544.
14. 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, http://doi.org/10.1016/j.bbr.2010.06.035.
15. Salomons, A.R., Kortleve, T., Reinders, N.R., Kirchhoff, S., Arndt, S.S., Ohl, F., 2010. Susceptibility of a potential animal model for pathological anxiety to chronic mild stress. Behav. Brain Res. 209 (2), 241-248, http://doi.org/10.1016/j.bbr.2010.01.050.
16. 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, http://doi.org/10.1016/j.bbr.2013.02.023.
17. Bothe, G.W.M., Bolivar, V.J., Vedder, M.J, Geistfeld, J.G., 2005. Behavioral differences among fourteen inbred mouse strains commonly used as disease models. Comp. Med. 55 (4), 326-334, PMID: 16158908.
18. Belzung, C., Griebel, G., 2001. Measuring normal and pathological anxiety-like behavior in mice: a review. Behav. Brain Res. 125 (1-2), 141-149, http://doi.org/10.1016/S0166-4328(01)00291-1.
19. Van der Goot, M. H., Boleij, H., van den Broek, J., Salomons, A. R., Arndt, S. S., van Lith, H. A., 2020. An individual based, multidimensional approach to identify emotional reactivity profiles in inbred mice. J. Neurosci. Meth. https://doi.org/10.1016/j.neumeth.2020.108810.
20. 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.
21. Garner, J.P., 2005. Stereotypies and other abnormal repetitive behaviors: Potential impact on validity, reliability, and replicability of scientific outcomes. ILAR Journal 46 (2), 106-117, http://doi.org/10.1093/ilar.46.2.106.
22. Tam, W. Y., Cheung, K-K., 2020. Phenotypic characteristics of commonly used mouse inbred strains. J. Mol. Med. Jul 25. https://doi.org/10.1007/s00109-020--1953-4.
23. Rogers, D. C., Jones, D. N. C., Nelson, P. R., Jones, C. M., Quilter, Ch. A., Robinson, T. L., Hagan, J. J., 1999. Use of SHIRPA and discriminant analysis to characterize marked differences in the behavioural phenotype of six inbred mouse strains. Behav. Brain Res. 105 (2), 207-217, https://doi.org/10.1016/s0166-4328(99)00072-8.
24. Bolivar, V. J., Caldarone, B. J., Reilly, A. A., Flaherty, L., 2000. Habituation of activity in an open field: A survey of inbred strains and F1 hybrids. Behav. Genet. 30, 285-293, https://doi.org/10.1023/A:1026545316455.

25. Griebel, G., Belzung, C., Perrault, G., Sanger, D. J., 2000. Differences in anxiety related behaviours and in sensitivity to diazepam in inbred and outbred strains of mice. Psychopharmacology, 148 (2), 164-170, https://doi.org/10.1007/s002130050038.
26. de Visser, L., van den Bos, R., Kuurman, W. W., Kas, M. J. H., Spruijt, B. M., 2006. Novel approach to the behavioural characterization of inbred mice: automated home cage observations. Genes Brain Behav. 5 (6), 458-466, https://doi.org/10.1111/j.1601-183X.2005.00181.x.
27. Ohl, F., 2003. Testing for anxiety. Clin. Neurosci. Res. 3 (4-5), 233-238, https://doi.org/10.1016/S1566-2772(03)00084-7.
28. Korte, S. M., 2001. Corticosteroids in relation to fear, anxiety, and psychopathology. Neurosci. Biobehav. Rev. 25 (2), 117-142, https://doi.org/10.1016/s0149-7634(01)00002-1.
29. Sgoifo, A., de Boer, S. F., Haller, J., Koolhaas, J. M., 1996. Individual differences in plasma catecholamine and corticosterone stress responses of wild-type rats. Physiol. Behav. 60 (6), 1403-1407, https://doi.org/10.1016/S0031-9384(96)00229-6.
30. Rougé-Pont, F., Deroche, V., Le Moal, M., Piazza, P. V., 1998. Individual differences in stress-induced dopamine release in the nucleus accumbens are influenced by corticosterone. Eur. J. Neurosci. 10 (12), 3903-3907, https://doi.org/10.1046/j.1460-9568.1998.00438.x.
31. Cockrem, J. F., 2013. Individual variation in glucocorticoid stress responses in animals. Gen. Comp. Endocrinol. 181, 45-58, https://doi.org/10.1016/j.ygcen.2012.11.025.
32. Jakovcevski, M., Schachner, M., Morellini, F., 2008. Individual variability in the stress response of C57BL/6J male mice correlates with trait anxiety. Genes Brain Behav. 7, 235-243, https://doi.org/10.1111/j.1601-183X.2007.00345.x
33. Jakovcevski, M., Schachner, M., Morellini, F., 2011. Susceptibility to the long-term anxiogenic effects of an acute stressor is mediated by the activation of the glucocorticoid receptors. Neuropharmacology 61 (8), 1297-1305, https://doi.org/10.1016/j.neuropharm.2011.07.034.
34. Nasca, C., Bigio, B., Zelli, D., Nicoletti, F., McEwen B. S. 2015. Mind the gap: glucocorticoids modulate hippocampal glutamate tone underlying individual differences in stress susceptibility. Mol. Psychiatry 20, 755-763, https://doi.org/10.1038/mp.2014.96.
35. Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lidster, K., MacCallum, C. J., Macleod, M., Petersen, O., Rawle, F., Reynolds, P., Rooney, K., Sena, E. S., Silberberg, S. D., Steckler, T., Würbel, H., 2020. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. BMC Vet Res. 16, 242, https://doi.org/10.1186/s12917-020-02451-y.
36. Percie du Sert N., Ahluwalia A., Alam S., Avey M. T., Baker M., Browne W. J., Clark A., Cuthill I. C., Dirnagl U., Emerson M., Garner P., Holgate S. T., Howells D. W., Hurst V., Karp N. A., Lazic S. E., Lidster K., MacCallum C. J., Macleod M., Pearl E. J., Petersen O. H., Rawle F., Reynolds P., Rooney K., Sena E. S., Silberberg S. D., Steckler T., Würbel H. 2020. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLos Biol. 18 (7), e3000411, https://doi.org/10.1371/journal.pbio.3000411.

37. 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.
38. Arndt, S.S., Laarakker, M.C., van Lith, H.A., van der Staay, F.J., Gieling, E., Salomons, A.R., van 't Klooster, J., Ohl, F. 2009. Individual housing of mice – impact on behavior and stress responses. Phys. Behav. 97 (3), 385-393, https://doi.org/10.1016/j.physbeh.2009.03.0008.
39. Festing, M.F.W., Baumans, V., Combes, R.D., Halder, M., Hendriksen, C.F.M., Howard, B.R., Lovell, D.P., Moore, G.J., Overend, P., Wilson, M.S. 1998. Reducing the of laboratory animals in biomedical research: problems and possible solutions. Altern. Lab. Anim. 26 (3), 283-301, PMID: 26042346.
40. Shaw, R., Festing, M.F.W., Peers, I., Furlong, L. 2002. Use of factorial designs to optimize animal experiments and reduce animal use. ILAR Journal 43(4), 223-232, https://doi.org/10.1093/ilar.43.4.223.
41. Buch, T., Moos, K., Ferreira, F.M., Fröhlich, H, Gebhard, C., Tresch, A. 2019. Benefits of a factorial design focussing on inclusion of female and male animals in one experiment. J. Mol. Med. 97, 871-877, https://doi.org/10.1007/s00109-019-01774-0.
42. 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.
43. Gouveia, K., Hurst, J., 2013. Reducing mouse anxiety during handling: effect of experience with handling tunnels. PLoS One 8 (6), e66401, https://doi.org/journal.pone.0066401.
44. Ohl, F., Sillaber, I., Binder, E., Keck, M.E., Holsboer, F. 2001. Differential analysis of behavior and diazepam-induced alterations in C57BL/6N and BALB/c mice using the modified hole board test. J. Psychiatr. Res. 35 (3), 147-154, https://doi.org/10.1016/S0022-3956(01)00017-6.
45. 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, http://doi.org/10.3791/52529.
46. Cicchetti, D. V., 2001. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsyc. 23 (5), 695-700, https://doi.org/10.1076/jcen.23.5.695.1249.
47. Honma S., Honma K.I., Shirakawa T., Hiroshige T., 1988. Rhythms in behaviors, body temperature and plasma corticosterone in SCN lesioned rats given methamphetamine. Physiol. Behav. 44 (2), 247–55, https://doi.org/10.1016/0031-9384(88)90146-1.
48. Dallman M. F., Bhatnagar S. Chronic stress and energy balance: role of the hypothalamo–pituitary–adrenal axis. In: McEwen B.S., Goodman H.M., editors. Handbook of physiology; section 7: the endocrine system; volume iv: coping with the environment: neural and endocrine mechanisms. Oxford University Press; New York, NY: 2001. pp. 179–210.
49. Dürschlag, M., Würbel, H., Stauffacher, M., Von Holst, D. 1996. Repeated blood collection in the laboratory mouse by tail incision – modification of an old technique. Phsyiol. Behav. 60 (6), 1565-1568. https://dx.doi.org/10.1016/S0031-9384(96)00307-1.

50. 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.
51. 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, http://doi.org/10.1016/j.physbeh.2010.10.019.
52. 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, http://doi.org/10.1016/j.jneumeth.2017.09.003.
53. 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, http://doi.org/10.1016/j.jneumeth.2011.01.019.
54. Cook, R. D. and Weisberg, S. (1982*). Residuals and Influence in Regression. New* York: Chapman and Hall. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/37076.
55. R Core Team, 2020. R: A language environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
56. Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., R Core Team, 2020. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1.-147, URL: https://CRAN.R-project.org/package=nlme.
57. Brooks, M. E., Kristensen K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., Bolker, B. M., 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modelling. R. J. 9 (2), 378-400.
58. 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.
59. 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. New York, NY: Springer, https://doi.org/10.1007/978-0-387-87458-6
60. Lenth, R., 2020. Emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.7, https://CRAN.R-project.org/package-emmeans.
61. Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. 62 (318), 626-633, https://doi.org/10.1080/01621459.1967.1048295.
62. Wahlsten, D., 2011. Sample size. In: Mouse behavioral testing: how to use mice in behavioral neuroscience, 1st edn., Academic Press, Elsevier Inc., London, U.K., pp 75-105.

63. Bate, S.T. and Clark, R.A. 2014. The Design and Statistical Analysis of Animal Experiments. Cambridge University Press, UK.
64. Genolini, C., Alacoque, X., Sentenac, M., Arnaud, C., 2015. Kml and kml3d: R-packages to cluster longitudinal data. J. Stat. Softw. 65(4), 1-34, http://www.jstatsoft.org/v65/i04/.
65. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.
66. Tibshirani, R., Walther, G., Hastie, T., 2002. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B. 63 (2), 411-423, https://doi.org/10.1111/1467-9868.00293.
67. Kryszczuk, K., Hurley, P., 2010. Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. In: El Gayar, N., Kittler, J., Roli, F. (eds). Multiple Classifier Systems. MCS 2010. Lecture Notes in Computer Science, vol 5997. Springer, Berlin, Heidelberg. https://doi.org/10.1007.
68. 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.
69. 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.
70. Ardayfio, P., Kim, K. 2006. Anxiogenic-like effect of chronic corticosterone in the light-dark emergence task in mice. Behav. Neurosci. 120 (2), 249-256, https://doi.org/10.1037/0735-7044.120.2.249.
71. 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, http://doi.org/10.1007/s10519-012-9572-8.
72. Ebner, K., Singewald, N., 2017. Individual differences in stress susceptibility and stress inhibitory mechanisms. Curr. Opin. Behav. Sci. 14, 65-64, https://doi.org/10.1016/j.cobeha.2016.11.016
73. Weger, M., Sandi, C. 2018. High anxiety trait: A vulnerable stress phenotype for stress-induced depression. Neurosci. Biobehav. Rev. 87, 27-37, https://doi.org/10.1016/j.neubiorev.2018.01.012.
74. Piazza, P.V., Maccari, S., Deminiere, J.M., LeMoal, M., Mormede, P., Simon, H., 1991. Corticosterone levels determine individual vulnerability to amphetamine self-administration. Proc. Natl. Acad. Sci. U S A, 88 (6), 2088-2092, https://doi.org/10.1073/pnas.88.6.2088.
75. 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.
76. Ducottet, C., Belzung, C., 2004. Behaviour in the Elevated-Plus-Maze predicts coping after subchronic mild stress in mice. Physiol. Behav. 81 (3), 417-426, https://doi.org/10.1016/j.physbeh.2004.01.013.
77. Cook, M.N., Bolivar, V.J., 2002. Behavioral differences among 129 substrains: Implication for knockout and transgenic mice. Behav. Neurosci. 116 (4), 600-611, https://doi.org/10.1037/0735-7044.116.4.600.
78. Pratte, M., Jamon, M., 2009. Detection of social approach in inbred mice. Behav. Brain Res. 203, 54-64, https://doi.org/10.1016/j.bbr.2009.04.011.
79. Kakihana, R., Moore, J.A. 1976. Circadian rhythm of corticosterone in mice: The effect of chronic consumption of alcohol. Phsychopharmacologia, 46, 301-305, https://10.1007/BF00421118.
80. Neuman, H., Debelius, J.W., Knight, R., Koren, O. 2015. Microbial endocrinology: the interplay between the microbiota and the endocrine system. FEMS Microbiol. Rev. 39 (4), 509-521, https://doi.org/10.1093/femsre/fuu010.
81. Jensen, T.L., Kiersgaard, M.K., Sorensen, D.B., Mikkelsen, L.F. 2013. Fasting of mice: a review. Lab. Anim. 47 (4), 225-240, https://doi.org/10.1177/0023677213501659.
82. Brinks. V., van der Mark, M., de Kloet, R., Oitzl, M., 2007. Emotion and cognition in high and low stress sensitive mouse strains: a combined neuroendocrine and behavioral study in BALB/c and C57BL/6 mice. Front. Behav. Neurosci. 1 (8), https://doi.org/10.3389/neuro.08.008.2007.
83. Sartori, S.B., Landgraf, R., Singewald, N., 2011. The clinical implications of mouse models of enhanced anxiety. Future Neurol. 6 (4), 531-571, https://doi.org/10.2217/fnl.11.34.
84. Lewejohann, L., Zipser, B., Sachser, N., 2011. „Personality" in laboratory mice used for biomedical research: A way of understanding variability? Dev. Psychobiol. 53 (6), 624-630, https://doi.org/10.1002/dev.20553.
85. Keshavarz, M., Krebs-Watson, R., Refki, P., Savriama, Y., Zhang, Y., Guenther, A., Brückl, T. M., Binder, E. B., Tautz, D., 2020. Natural copy number variation differences of tandemly repeated small nucleolar RNAs in the Prader-Willi syndrome genomic region regulate individual behavioral responses in mammals. bioRxiv 476010, https://doi.org/10.1101/476010.
86. 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, http://doi.org/10.1126/science.1235294.
87. Labots. M., Zheng, X., Moattari, G., Ohl, F., van Lith, H.A., 2016. Effects of light regime and substrain on behavioral profiles of male C57BL/6 mice in three tests of unconditioned anxiety. J. Neurogenet. 30 (4), 306-315, https://doi.org/10.1080/01677063.2016.1249868.
88. Braun, E., Geurten, B., Egelhaaf, M., 2010. Identifying prototypical components in behaviour using clustering algorithms. PLoS ONE 5 (2), e9361, https://doi.org/10.1371/journal.pone.0009361.
89. Kiselev, V.Y., Andrews, T.S., Hemberg, M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. 20, 273-282, https://doi.org/10.1038/s41576-018-0088-9.

90. Bouwknecht, J.A., Olivier, B., Paylor, R.E. 2007. The stress-induced hyperthermia paradigm as a physiological animal model for anxiety: a review of pharmacological and genetic studies in the mouse. Neurosci. Biobehav. Rev. 31 (1), 41-59, https://doi.org/10.1016/j.neubiorev.2006.02.002.
91. Bouwknecht, J.A., Paylor, R. 2008. Pitfalls in the interpretation of genetic and pharmacological effects on anxiety-like behaviour in rodents. Behav Pharmacol. 19(5-6), 385-402, https://doi.org/10.1097/FBP.0b013e32830c3658.
92. Festing, M.F.W., 2014. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. ILAR Journal 55 (3), 399-404, http://doi.org/10.1093/ilar/ilu036.
93. Richter, S.H., Garner, J.P., Auer, C., Kunert, J., Wuerbel, H. 2010. Systematic variation improves reproducibility of animal experiments. Nat. Meth. 7, 167-168, https://doi.org/10.1038/nmeth0310-167.
94. Richter S.H. Systematic heterogenization for better reproducibility in animal experimentation. 2017. Lab. Anim. 46, 343-349, https://doi.org/10.1038/laban.1330.
95. Kafkafi, N., Agassi, J., Chesler, E.J. 2018. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci. Biobehav. Rev. 87, 218-232, https://doi.org/10.1016/j.neubiorev.2018.01.003.
96. Bello, N.M., Renter, D.G. 2018. Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. J. Dairy Sci. 101, 5679-5701, https://doi.org/10.3168/jds.2017-13978.
97. Bohlen, M., Hayes, E. R., Bohlen, B., Bailoo, B. D., Crabbe, J. C., Wahlsten, D., 2014. Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. Behav. Brain Res. 217, 46-54, https://doi.org/10.1016/j.bbr.2014.06.017.
98. Wahlsten, D., Metten, P., Philips, T. J., Boehm, S. L., , Burkhart-Kasch, S., Dorow, J., 2003. Different data from different labs: lessons from studies of gene-environment interaction. J. Neurobiol. 54, 283-311, https://doi.org/10.1002/neu.10173.
99. Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L., Mogil, J. S., 2002. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. Neurosci. Biobehav. Rev. 26, 907-923. https://doi.org/10.1016/s0149-7634(02)00103-3.
100. 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, 1101-1102. https://doi.org/10.1038/nn1102-1101.
101. Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., Wieskopf, J. S., Acland, E. L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J. C. S., McPhail, M., Delaney, A., Wigerblad, G., Schumann, A. P., Quinn, T., Frasnelli, J., Svensson, C. I., Sternberg, W. F., Mogil, J. S. 2014. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat. Methods 11, 629-632, https://doi.org/10.1038/nmeth.2935.

102. Van Driel, K. S., Talling, J. C., 2005. Familiarity increases consistency in animal tests. Behav. Brain Res. 159, 243-245, https://doi.org/10.1016/j.bbr.2004.11.005.
103. Richter, S. H. 2020. Automated home-cage testing as a tool to improve reproducibility of behavioral research? Front. Neurosci. 14, 383, https://doi.org/10.3389/fnins.2020.00383.
104. Kaufman, A. B., Rosenthal, R., 2009. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Anim. Behav. 78 (6), 1478-1491, https://doi.org/j.anbehav.2009.09.014.
105. Mogil, J.S. 2017. Laboratory environment factors and pain behavior: the relevance of unknown unknowns to reproducibility and translation. Lab Animal 46, 136-141, https://doi.org/10.1038/laban.1223.
106. Niemczura, A.C., Grimsley, J.M., Kim, C., Alkhawaga, A., Poth, A., Carvalho, A., Wenstrup, J.J. 2020. Physiological and behavioral response to vocalization playback in mice. Front. Behav. Neurosci. 14, 155, https://doi.org/10.3389/fnbeh.2020.00155.
107. 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.
108. Donner, N. C., Lowry, C. A., 2013. Sex differences in anxiety and emotional behavior. Eur. J. Physiol., 465, 601-626, https://doi.org/10.1007/s00424-013-1271-7.
109. 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.

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

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

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

Motivational system/Behavioral dimension	Behavioral variable	Directionality z-score[1]
Anxiety related behavior		
Avoidance behavior	Total number of board entries	-z
	Latency until first board entry	z
	Percentage of time spent on the board	-z
Risk assessment	Total number of stretched attends	z
	Latency until first stretched attend	-z
Arousal	Total number of self-groomings	z
	Latency until the first self-grooming	-z
	Percentage of time self-grooming	z
	Total number of boli	z
	Latency until first boli is produced	-z
Activity		
Exploration	Total number of rearings in the box	z
	Latency until first rearing in the box	-z
	Total number of rearings on the board	z
	Latency until first rearing on the board	-z
	Total number of hole explorations	z
	Latency until first hole exploration	-z
	Total number of hole visits	z
	Latency until first hole visit	-z
Locomotion	Total number of line crossings	z
	Latency until first line crossing	-z

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

Table S3. Strain differences: Effects of different explanatory variables for each behavioral dimension/pCORT. Behavioral dimensions were analyzed with generalized linear mixed models (GLMMs) using a 3 (strain) x 2 (experimenter) x 4 (trials) mixed factorial design. Strain, experimenter, trial and their interactions were included as fixed predictors. Mouse identity (ID), mouse slope (trial nested in ID), batch and testorder were included as random factors.

pCORT: linear mixed model (LMM) using a 3 (strain) x 4 (technician) x 3 (sampling moment) mixed factorial design. Strain, sampling moment and bio-technician were included as fixed predictors, including their interactions. Day of test was included as fixed covariate. Mouse identity, mouse slope (sampling moment nested in ID), batch, testorder were included as random factors. Significant effects are highlighted in bold.

Dimension	Explanatory variables	F	df	P
(a)Avoidance				
	Strain S	17.44	2, 111	**< 0.0001**
	Trial T	11.04	3, 329	**< 0.0001**
	Experimenter E	8.61	1, 111	**0.0041**
	S*T	8.51	6, 329	**<0.0001**
	S*E	1.14	2, 111	0.3236
	T*E	0.70	3, 329	0.5529
	S*T*E	2.12	6, 329	0.0507
	Mouse identity[r]	117.68[b]	14	**< 0.0001**
	Mouse slope[r]	21.01[b]	16	**< 0.0001**
	batch[r]	0.25[b]	19	0.9686
	testorder[r]	0.00[b]	22	1.000
(b)Risk assessment				
glmmTMB				
	Strain S[a]	186.53[b]	2	**<.0001**
	Trial T[a]	678.71[b]	3	**<.0001**
	Experimenter E[a]	3.59[b]	1	0.0580
	S*T[a]	175.72[b]	6	**<.0001**
	S*E[a]	4.56[b]	2	0.1022
	T*E[a]	23.20[b]	3	**0.0001**
	S*T*E[a]	12.03[b]	6	0.0613
	Mouse identity[r]	14.95[b]	15	**0.0001**
	Mouse slope[r]	1.58[b]	17	0.4535
	batch[r]	-*	-	-
	testorder[r]	-	-	-
(c)Arousal				
	Strain S	3.49	2, 111	**0.0339**
	Trial T	20.66	3, 329	**<.0001**
	Experimenter E	9.15	1, 111	**0.0031**
	S*T	3.57	6, 329	**0.0019**
	S*E	0.75	2, 111	0.4732
	T*E	2.45	3, 329	0.0631
	S*T*E	1.79	6, 329	0.1011
	Mouse identity[r]	22.82	14	**<0.0001**
	Mouse slope[r]	13.08	16	**0.0014**
	batch[r]	0.87	19	0.8328
	testorder[r]	0.01	22	0.9998
(d)Exploration				
	Strain S	38.15	2, 111	**< 0.0001**
	Trial T	79.00	3, 329	**< 0.0001**
	Experimenter E	114.95	1, 111	**< 0.0001**
	S*T	3.73	6, 329	**0.0113**
	S*E	0.34	2, 111	0.7117
	T*E	5.08	3, 329	**0.0019**
	S*T*E	0.84	6, 329	0.5367
	Mouse identity[r]	94.83[b]	14	**< 0.0001**
	Mouse slope[r]	39.59[b]	16	**< 0.0001**
	batch[r]	0.38[b]	19	0.9445
	testorder[r]	0.05[b]	22	0.9968
(e)Locomotion				
rank transformed				
	Strain S	183.02	2, 111	**< 0.0001**
	Trial T	18.91	3, 329	**< 0.0001**
	Experimenter E	5.26	1, 111	**0.0237**
	S*T	20.94	6, 329	**< 0.0001**
	S*E	3.50	2, 111	**0.0336**
	T*E	2.36	3, 329	0.0711
	S*T*E	0.76	6, 329	0.5977
	Mouse identity[r]	111.56[b]	14	**< 0.0001**
	Mouse slope[r]	27.61[b]	16	**< 0.0001**
	batch[r]	1.41[b]	19	0.7026
	testorder[r]	na	Na	Na
(f) pCORT				
	Strain S	21.81	2, 321	**< 0.0001**
	Sampling time (T)	263.98	2, 321	**< 0.0001**
	Biotechnician (B)	0.31	3, 321	0.8173
	S*T	4.67	4, 321	**0.0011**
	S*B	Na	Na	na
	T*B	Na	Na	na
	S*T*B	Na	Na	na
	Day of test (D)	0.12	1, 321	0.7231
	Mouse identity[r]	1.87	15	0.1718
	Mouse slope[r]	0.00	17	1.000
	batch[r]	Na	Na	na
	testorder[r]	Na	Na	na

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

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

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

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

Dimension		Estimate ± SEM	$t_{(df)}$	*P*	Cohens *d*	Lower and upper limits of 95%CI
(a)Avoidance						
Trial 1	C vs B6N	0.648 ± 0.178	$3.646_{(111)}$	**0.0004**	2.372	1.045, 3.699
	C vs 129S2	1.529 ± 0.162	$9.422_{(111)}$	**< 0.0001**	5.595	4.203, 6.987
	B6N vs 129S2	0.881 ± 0.152	$5.815_{(111)}$	**< 0.0001**	3.223	2.044, 4.402
Trial 2	C vs B6N	0.394 ± 0.139	$2.820_{(111)}$	**0.0057**	1.439	0.410, 2.468
	C vs 129S2	0.729 ± 0.148	$4.910_{(111)}$	**< 0.0001**	2.666	1.533, 3.799
	B6N vs 129S2	0.335 ± 0.142	$2.362_{(111)}$	0.0199	1.227	0.185, 2.270
Trial 3	C vs B6N	0.060 ± 0.137	$0.442_{(111)}$	0.6595	0.221	-0.772, 1.215
	C vs 129S2	0.333 ± 0.150	$2.222_{(111)}$	0.0283	1.217	0.119, 2.315
	B6N vs 129S2	0.272 ± 0.156	$1.746_{(111)}$	0.0836	0.996	-0.142, 2.133
Trial 4	C vs B6N	-0.031 ± 0.137	$-0.242_{(111)}$	0.8093	-0.121	-1.113, 0.871
	C vs 129S2	0.167 ± 0.155	$1.080_{(111)}$	0.2825	0.613	-0.514, 1.740
	B6N vs 129S2	0.201 ± 0.164	$1.227_{(111)}$	0.2224	0.734	-0.455, 1.923
(b)Risk assessment						
Trial 1	C vs B6N	1.561 ± 0.128	$12.211_{(432)}$	**< 0.0001**	4.195	3.463, 4.926
	C vs 129S2	0.340 ± 0.123	$2.754_{(432)}$	0.0061	0.914	0.259, 1.569
	B6N vs 129S2	-1.221 ± 0.120	$-10.181_{(432)}$	**< 0.0001**	-3.281	-3.951, -2.160
Trial 2	C vs B6N	0.991 ± 0.095	$10.389_{(432)}$	**< 0.0001**	2.661	2.127, 3.195
	C vs 129S2	0.028 ± 0.116	$0.244_{(432)}$	0.8073	0.076	-0.535, 0.687
	B6N vs 129S2	-0.962 ± 0.096	$-10.047_{(432)}$	**< 0.0001**	-2.586	-3.120, 02.051
Trial 3	C vs B6N	0.250 ± 0.078	$3.215_{(432)}$	**0.0014**	0.671	0.259, 1.085
	C vs 129S2	-0.109 ± 0.109	$-0.999_{(432)}$	0.3185	-0.291	-0.866, 0.283
	B6N vs 129S2	-0.359 ± 0.087	$-4.129_{(432)}$	**< 0.0001**	-0.963	-1.426, -0.500
Trial 4	C vs B6N	0.156 ± 0.067	$2.360_{(432)}$	0.0187	0.419	0.069, 0.769
	C vs 129S2	0.045 ± 0.104	$-0.437_{(432)}$	0.6624	-0.122	-0.670, 0.426
	B6N vs 129S2	-0.201 ± 0.084	$-2.386_{(432)}$	0.0175	-0.541	-0.988, -0.094
(c)Arousal						
Trial 1	C vs B6N	-0.151 ± 0.104	$-1.460_{(111)}$	0.1472	-0.347	-0.819, 0.126
	C vs 129S2	0.255 ± 0.103	$2.472_{(111)}$	0.0149	0.584	0.109, 1.059
	B6N vs 129S2	0.406 ± 0.106	$3.847_{(111)}$	**0.0002**	0.931	0.436, 1.426
Trial 2	C vs B6N	0.226 ± 0.111	$2.042_{(111)}$	0.0435	0.517	0.011, 1.024
	C vs 129S2	0.211 ± 0.111	$1.908_{(111)}$	0.0590	0.483	-0.028, 0.989
	B6N vs 129S2	-0.015 ± 0.113	$-0.133_{(111)}$	0.8941	-0.034	-0.546, 0.477
Trial 3	C vs B6N	0.058 ± 0.123	$0.471_{(111)}$	0.6389	0.132	-0.424, 0.689
	C vs 129S2	0.014 ± 0.121	$0.118_{(111)}$	0.9062	0.032	-0.514, 0.580
	B6N vs 129S2	-0.434 ± 0.124	$-0.350_{(111)}$	0.7269	-0.099	-0.663, 0.464
Trial 4	C vs B6N	-0.196 ± 0.135	$-1.458_{(111)}$	0.1477	-0.450	-1.064, 0.165
	C vs 129S2	-0.125 ± 0.134	$-0.933_{(111)}$	0.3528	-0.286	-0.894, 0.322
	B6N vs 129S2	0.072 ± 0.136	$0.525_{(111)}$	0.6005	0.164	-0.455, 0.784
(d)Exploration						
Trial 1	C vs B6N	-0.437 ± 0.071	$-6.114_{(111)}$	**< 0.0001**	-1.451	-1.960, -0.943
	C vs 129S2	0.017 ± 0.070	$0.247_{(111)}$	0.8050	0.057	-0.401, 0.515
	B6N vs 129S2	0.454 ± 0.071	$6.416_{(111)}$	**< 0.0001**	1.509	1.001, 2.016
Trial 2	C vs B6N	-0.531 ± 0.090	$-5.915_{(111)}$	**< 0.0001**	-1.764	-2.401, -1.128
	C vs 129S2	-0.071 ± 0.096	$-0.742_{(111)}$	0.4599	-0.237	-0.870, 0.396
	B6N vs 129S2	0.460 ± 0.101	$4.554_{(111)}$	**< 0.0001**	1.528	0.833, 2.223
Trial 3	C vs B6N	-0.399 ± 0.107	$-3.733_{(111)}$	**0.0003**	-1.324	-2.049, -0.599
	C vs 129S2	0.175 ± 0.119	$1.468_{(111)}$	0.1450	0.580	-0.207, 1.367
	B6N vs 129S2	0.573 ± 0.124	$4.626_{(111)}$	**< 0.0001**	1.904	1.050, 2.758
Trial 4	C vs B6N	-0.169 ± 0.093	$-1.807_{(111)}$	0.0736	-0.561	-1.180, 0.058
	B6N vs 129S2	0.428 ± 0.119	$3.587_{(111)}$	**0.0005**	1.423	0.614, 2.231
(e)Locomotion						
Trial 1	C vs B6N	-241.65 ± 17.2	$-14.065_{(111)}$	**< 0.0001**	-2.763	-3.299, -2.228
	C vs 129S2	85.82 ± 20.2	$4.240_{(111)}$	**< 0.0001**	0.981	0.505, 1.458
	B6N vs 129S2	327.47 ±17.4	$18.809_{(111)}$	**< 0.0001**	3.745	3.109, 4.380
Trial 2	C vs B6N	-116.75 ± 18.1	$-6.456_{(111)}$	**< 0.0001**	-1.335	-1.782, -0.889
	C vs 129S2	101.16 ± 21.0	$4.814_{(111)}$	**< 0.0001**	1.157	0.656, 1.657
	B6N vs 129S2	217.91 ± 18.3	$11.904_{(111)}$	**< 0.0001**	2.492	1.961, 3.023
Trial 3	C vs B6N	-68.03 ± 19.5	$-3.492_{(111)}$	**0.0007**	-0.778	-1.231, -0.325
	C vs 129S2	131.53 ± 22.1	$5.960_{(111)}$	**< 0.0001**	1.504	0.965, 2.043
	B6N vs 129S2	199.55 ± 19.6	$10.170_{(111)}$	**< 0.0001**	2.282	1.744, 2.820
Trial 4	C vs B6N	-6.82 ± 21.1	$-0.324_{(111)}$	0.7467	-0.078	-0.555, 0.399
	C vs 129S2	159.23 ± 23.5	$6.786_{(111)}$	**< 0.0001**	1.824	1.239, 2.410
	B6N vs 129S2	166.35 ± 21.2	$7.834_{(111)}$	**< 0.0001**	1.902	1.359, 2.446
(f)pCORT						
Time 1	C vs B6N	0.13 ± 0.06	$2..168_{(79)}$	0.0391	0.164	0.007, 0.321
	C vs 129S2	-0.02 ± 0.06	$-0.437_{(78)}$	0.6656	-0.031	-0.179, 0.116
	B6N vs 129S2	-0.15 ± 0.06	$-2.436_{(77)}$	0.0217	-0.196	-0.363, -0.029
Time 2	C vs B6N	-0.10 ± 0.22	$-0.441_{(79)}$	0.6629	-0.123	-0.694, 0.448
	C vs 129S2	-0.87 ± 0.32	$-2.750_{(78)}$	0.0104	-1.102	-1.937, -0.267
	B6N vs 129S2	-0.78 ± 0.27	$-2.857_{(77)}$	**0.0081**	-0.979	-1.694, -0.264

Rank transformed

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

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

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

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

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

	Mouse identity[r]	184.36 [b]	11	**< 0.0001**
	Mouse slope[r]	Na	na	na
(f)pCORT				
	Cluster C	2.73	1, 115	0.2429
	Time T	224.8	2, 213	**< 0.0001**
	C*T	2.11	2, 213	0.0880
	Mouse identity[r]	4.51 [b]	8	**0.0336**
	Mouse slope[r]	0.00 [b]	10	1.0000

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

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

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

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

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

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

Results section		GLMM effect	Post hoc comparisons/ contrasts	γ	Adjusted α
Strain analyses	Behavior	Strain	C vs B6; C vs 129S2; B6 vs 129S2	2	0.025321
	Behavior	Trial	Trial 1 vs Trial 4	1	0.05
	Behavior	Strain x Trial (T)	C-T1 vs C-T4; C-T1 vs B6-T1; C-T1 vs 129S2-T1	3	0.01692
			B6-T1 vs B6-T4; C-T1 vs B6-T1; B6-T1 vs 129S2-T1	3	0.01692
			129S2-T1 vs 129S2-T4; C-T1 vs 129S2-T1; B6-T1 vs 129S2-T1	3	0.01692
			C-T2 vs B6-T2; C-T2 vs 129S2-T2; B6-T2 vs 129S2-T2	2	0.025321
			C-T3 vs B6-T3; C-T3 vs 129S2-T3; B6-T3 vs 129S2-T3	2	0.025321
	Behavior	Experimenter (E) x Trial (T)	E1-T1 vs E1-T4; E1-T1 vs E2-T1; E1-T4 vs E2-T4	2	0.025321
			E2-T1 vs E2-T4; E1-T1 vs E2-T1; E1-T4 vs E2-T4	2	0.025321
			E1-T2 vs E2-T2	1	0.05
			E1-T3 vs E2-T3	1	0.05
		Experimenter (E) x Strain	E1-C vs E2-C; E1-B6 vs E2-B6; E1-129S2 vs E2 – 129S2	1	0.05
	Corticosterone	Strain x Sampling moment (S)	**C-S1** vs C-S2; **C-S1** vs C-S3; C-S2 vs C-S3; B6-S1 vs B6-S2; B6-S1 vs B6-S3; B6-S2 vs B6-S3; ; 129S2-S1 vs 129S2-S2; 129S2-S1 vs 129S2-S3; 129S2-S2 vs 129S2-S3;	4	0.012741
			C-S1 vs B6-S1; **C-S1** vs 129S2-S1; B6-S1 vs 129-S1; C-S2 vs B6-S2; C-S2 vs 129S2-S2; B6-S2 vs 129-S2; C-S3 vs B6-S3; C-S1 vs 129S2-S3; B6-S3 vs 129-S3;		
Cluster analyses	Behavior	Cluster (A/B) x Trial (T)	A-T1 vs A-T4; B-T1 vs B-T4; A-T1 vs B-T1; A-T4 vs B-T4	2	0.025321
			A-T2 vs B-T2; A-T3 vs B-T3	1	0.05
	Corticosterone	Trial (T)	T1 vs T2; T1 vs T3; T2 vs T3	2	0.025321

Table S11. Cluster (A/B) differences within strains C, B6N and 129S2: Effects of different explanatory variables for each behavioral dimension/pCORT. Behavioral dimensions were analyzed with generalized linear mixed models (GLMMs) using a 2 (cluster) x 4 (trials) mixed factorial design. Cluster, trial and their interactions were included as fixed predictors. Mouse identity (ID)was included as random factor.

pCORT: linear mixed model (LMM) using a 2 (cluster) x 3 (sampling moment) mixed factorial design. Cluster, sampling moment were included as fixed predictors, including their interaction. Mouse identity was included as random factor. Significant effects ($P < 0.05$) are highlighted in bold.

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

Exploration	Cluster C	0.02	1, 37	0.8909
	Trial T	38.04	3, 109	**< 0.0001**
	C * T	7.11	3, 109	**0.0002**
Locomotion	Cluster C	6.30	1, 37	**0.0165**
	Trial T	63.36	3, 109	**< 0.0001**
	C * T	0.27	3, 109	0.8473
Corticosterone	Cluster C	0.18	1, 37	0.6689
	Trial T	154.34	2,67	**< 0.0001**
	C * T	2.89	2, 67	0.0627
Strain: 129S2				
Avoidance	Cluster C	1.06	1, 36	0.3101
	Trial T	0.88	3, 106	0.4533
	C * T	21.53	3, 106	**< 0.0001**
Arousal	Cluster C	1.32	1, 36	0.2585
	Trial T	17.89	3, 106	**<0.0001**
	C * T	0.08	3, 106	0.9719
Risk assessment (*glmmTMB*)	Cluster C	1.04	1, 36	0.3071
	Trial T	553.84	3, 106	**< 0.0001**
	C * T	4.69	3, 106	0.1958
Exploration	Cluster C	0.67	1, 36	0.4162
	Trial T	13.10	3, 106	**< 0.0001**
	C * T	7.51	3, 106	**0.0001**
Locomotion	Cluster C	1.43	1, 36	0.2396
	Trial T	0.68	3, 106	0.5673
	C * T	0.32	3, 106	0.8074
Corticosterone	Cluster C	0.34	1, 36	0.2599
	Time T	143.29	2, 65	**< 0.0001**
	C * T	0.41	2, 65	0.8197

Table S12. *Post hoc* comparisons of the estimated marginal means between and/or within clusters, for each strain (C/B6N/129S2) separately.

Within cluster *post hoc* comparisons between trials 1 and 4 (behavioral dimensions) or sampling moments (pCORT): Behavioral dimensions: adjusted α = 0.025321 for comparison between trial 1 and 4. pCORT: adjusted α = 0.025321 for comparison between sampling moments.

Post hoc between cluster comparisons on each trial: Behavioral dimensions: (adjusted α = 0.025321 for trials 1 and 4, α = 0.05 for trials 2 and 3). Significant contrasts are highlighted in bold.

Strain: C		**Estimate ± SEM**	$t_{(df)}$	**P**
Avoidance				
Cluster A	Trial 1 vs 4	1.241 ± 0.128	$9.661_{(114)}$	**< 0.0001**
Cluster B	Trial 1 vs 4	-0.236 ± 0.279	$-0.847_{(114)}$	0.3985
Trial 1				
	A vs B	0.944 ± 0.323	$2.919_{(38)}$	**0.0059**
Trial 2				
	A vs B	0.169 ± 0.245	$0.692_{(38)}$	0.4934
Trial 3				
	A vs B	-0.275 ± 0.221	$-1.245_{(38)}$	0.2208
Trial 4				
	A vs B	-0.534 ± 0.220	$-2.424_{(38)}$	**0.0202**
Risk assessment				
Trial main				
	Trial 1 vs 4	2.316 ± 0.151	$15.267_{(146)}$	**< 0.0001**
Arousal				
Cluster A				
	Trial 1 vs 4	-0.234 ± 0.099	$-2.376_{(114)}$	**0.0192**
Cluster B				
	Trial 1 vs 4	-0.854 ± 0.214	$-3.992_{(114)}$	**0.0001**
Trial 1				
	A vs B	0.159 ± 0.185	$0.860_{(38)}$	0.3951
Trial 2				
	A vs B	-0.036 ± 0.185	$-0.196_{(38)}$	0.8455
Trial 3				
	A vs B	-0.659 ± 0.185	$-3.556_{(38)}$	**0.0010**
Trial 4				
	A vs B	-0.460 ± 0.185	$-2.486_{(38)}$	**0.0174**

Exploration				
Trial main				
	Trial 1 vs 4	-0.685 ± 0.080	$-8.597_{(114)}$	**< 0.0001**
Locomotion				
Trial main				
	Trial 1 vs 4	-0.414 ± 0.108	$-3.807_{(114)}$	**0.0002**
Corticosterone				
Time main				
	Time 1 vs 2	-1.298 ± 0.203	$-6.394_{(73)}$	**< 0.0001**
	Time 1 vs 3	-0.507 ± 0.177	$-2.863_{(73)}$	**0.0055**
	Time 2 vs 3	0.791 ± 0.160	$4.938_{(73)}$	< 0.0001
Strain: B6N		Estimate ± SEM	$t_{(df)}$	P
Avoidance				
Cluster A				
	Trial 1 vs 4	0.644 ± 0.116	$5.573_{(109)}$	< 0.0001
Cluster B				
	Trial 1 vs 4	-0.307 ± 0.127	$-2.383_{(109)}$	**0.0189**
Trial 1				
	A vs B	0.714 ± 0.232	$3.079_{(37)}$	**0.0039**
Trial 2				
	A vs B	0.234 ± 0.231	$1.015_{(37)}$	0.3166
Trial 3				
	A vs B	-0.209 ± 0.231	$-0.905_{(37)}$	0.3715
Trial 4				
	A vs B	-0.233 ± 0.231	$-1.011_{(37)}$	0.3187
Risk assessment				
Trial main				
	Trial 1 vs 4	0.805 ± 0.072	$11.183_{(151)}$	**< 0.0001**
Arousal				
Cluster A				
	Trial 1 vs 4	-0.118 ± 0.134	$-0.885_{(109)}$	0.3779
Cluster B				
	Trial 1 vs 4	-0.771 ± 0.190	$-4.054_{(109)}$	0.0001
Trial 1				
	A vs B	0.180 ± 0.174	$1.038_{(37)}$	0.3060
Trial 2				
	A vs B	0.040 ± 0.171	$0.237_{(37)}$	0.8143

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

Trial 4				
	A vs B	-1.070 ± 0.265	$-4.038_{(36)}$	**0.0003**
Risk assessment				
Trial main				
	Trial 1 vs 4	1.821 ± 0.093	$19.470_{(144)}$	**< 0.0001**
Arousal				
Trial main				
	Trial 1 vs 4	-0.708 ± 0.120	$-5.920_{(106)}$	**< 0.0001**
Exploration				
Cluster A				
	Trial 1 vs 4	-1.172 ± 0.178	$-6.575_{(106)}$	**< 0.0001**
Cluster B				
	Trial 1 vs 4	-0.333 ± 0.100	$-3.319_{(106)}$	**0.0012**
Trial 1				
	A vs B	-0.250 ± 0.230	$-1.090_{(36)}$	0.2827
Trial 2				
	A vs B	-0.095 ± 0.230	$-0.412_{(36)}$	0.6829
Trial 3				
	A vs B	0.395 ± 0.229	$1.722_{(36)}$	0.0937
Trial 4				
	A vs B	0.588 ± 0.229	$2.566_{(36)}$	**0.0146**
Corticosterone				
Time main				
	Time 1 vs 2	-1.813 ± 0.150	$-12.074_{(65)}$	**< 0.0001**
	Time 1 vs 3	-0.273 ± 0.190	$-1.437_{(65)}$	0.1556
	Time 2 vs 3	1.540 ± 0.162	$9.532_{(65)}$	**< 0.0001**

Chapter 4

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

PLoS ONE, 2021, 16(8), e0255521.

Marloes H. van der Goot[1,4], Marieke Kooij[1], Suzanne Stolte[1], Annemarie Baars[1], Saskia S. Arndt[1,3], Hein A. van Lith[1,2,3]

[1] *Section Animals in Science and Society, Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands*
[2] *Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands*
[3] *Both authors share senior authorship*

Abstract

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

1. Introduction

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

Inter-individual variability however is often not actively accounted for in the design of animal experiments [3]. Traditionally, this type of variation was regarded as part of a larger source of unwanted noise, falling within the same category as other sources of extraneous noise (i.e. measurement error) and unanticipated environmental effects [3]. In contrast to these other sources however, inter-individual variability has been shown to be relatively robust to standardization efforts, distinguishing it from mere noise [2]. Therefore, an increasing body of research focuses on the identification of methods that consider inter-individual differences by systematically incorporating this variation in experimental design and statistical analysis [3][5][6][7][8][9][10][11][12].

The importance of incorporating inter-individual variability in the design of animal experiments has become especially acknowledged in animal models of behavioral dysfunction (e.g., anxiety, depression [5][13][14], post-traumatic stress disorder, [15][16] but also addiction [17] etc.). In humans, the susceptibility between individuals to develop a particular disorder, as well as the response to treatment is known to vary substantially between individuals [13]. Considering this variability may therefore not only improve reproducibility between studies, it may also contribute to an improved understanding of the mechanisms that underlie such inter-individual variability in human patients [5][13]. In this field, several strategies exist to incorporate inter-individual variability as a variable [17][18]. The most prominent strategy considers inter-individual variability between subpopulations within an experimental pool (mostly outbred stocks of rats and mice) by separating experimental animals whose expression of a

particular trait (i.e. anxiety, activity) lies on opposing ends of a phenotypic distribution, for example by means of a median, tertiary, quartile split [5][17][18]. Interestingly, this use of selection strategies has indirectly demonstrated how existing subpopulations within an experimental pool may mask the detection of overall group effects, thereby providing examples of how inter-individual variability may confound experimental outcomes [14]. Barbelivien et al. [19], for example identified five sub-populations of Long Evans outbred rats that were characterized by differential levels of baseline impulsive choice behavior. Subsequent administration of d-amphetamine only affected impulse choice behavior when these baseline differences were accounted for, while no effect was found when all animals were pooled in the analysis.

A fundamental principle of good design of animal experiments is that all variables should be controlled except that due to treatment and that all treatment and control groups should be identical, with minimal within-group variability [20][21]. Following these principles, accounting for inter-individual variability in the composition of experimental groups should result in better matched individuals regarding control and treatment, thereby improving the experimental design and the quality of the results. To our knowledge however, the extent to which active incorporation of inter-individual differences in the composition of experimental groups affects the outcome of preclinical animal experiments, has never been directly tested. In this study, we therefore compared the outcomes of an experimental design in which this inter-individual variability was accounted for, to a design in which this variability was not accounted for, to empirically assess to what extent active incorporation of inter-individual variability indeed alters the interpretation of a standard pharmacological experiment when evaluating the effects of an anxiolytic compound on anxiety-related behavior. To do so we defined the behavioral phenotype of experimental animals on an individual level in a pre-experimental period, and subsequently incorporated this information in the design and statistical analysis of our study.

A priori identification of subgroups within an experimental pool is also common in the aforementioned selection strategies. A disadvantage of these selection strategies however is that in the majority of these studies inter-individual variability is established by means of an artificially predetermined quantile (but see [22][23] for exceptions). Median split strategies may lead to a loss in resolution and power as every value above and below the mean is considered equal, regardless of its position on the phenotypic distribution [24].

Furthermore, strategies considering upper and lower quantiles only include the outer ends of a phenotypic distribution, rather than the entire study population [25]. These criteria contrast with the generally accepted conceptualization of human psychopathology as a continuum [26], which warrants the exploration of subgroups across the entire study population on the basis of variability in the data itself, rather than a predefined criterion [14]. In the present study we therefore used a data-driven clustering approach to identify meaningful subpopulations within our experimental pool (see below). Furthermore, instead of outbred stocks, we assessed inter-individual variability in mouse inbred strains. As outlined above, inter-individual differences in spontaneous behavior have been repeatedly demonstrated within mouse inbred strains, and have been found to be consistent over time within individuals [27][28][29]. In fact, inbred strains of mice have been demonstrated to be just as variable as outbred stocks of this species [30].

We expanded on a series of previous studies in which we found that our phenotype of interest, behavioral habituation of anxiety-related responses, may differ within BALB/c, C57BL/6 and various 129 sub-strains [31][32]. These studies measured behavioral habituation as the change in anxiety and activity related behavior after repeated exposure to a mild aversive stimulus (the modified Hole Board (mHB)). Anxiety is typically regarded as a complex behavioral construct that is expressed by both anxiety-related and activity behaviors [33][34][35]. Multivariate cluster analyses on the resulting individual response trajectories identified two clusters in which mice grouped together across both anxiety and activity related dimensions: individual response types. These response types were of differential adaptive value, and were displayed by individuals of all three strains.

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

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

2. Materials and Methods

2.1. Ethical statement

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

2.2. Animals and housing

This study tested naïve males of three mouse inbred strains: BALB/cAnNCrl (hereafter C, *n* = 59, white (albino), strain code = 028), C57BL/6NCrl (B6N, *n* = 60, black, strain code = 027) and 129S2/SvPasCrl (129S2, *n* = 60, agouti, strain code = 287). One additional C mouse died due to reasons unrelated to the study and was not tested. Furthermore, an additional number of 15 naïve males (*n* = 5/strain) were used to establish the required dose of pharmacological treatment in a pilot study. One B6N mouse died due to reasons unrelated to the pilot study and was not tested. The total number of animals used in the present study amounted to 179 + 14 = 193. The sample size was determined using the software by Lenth (www.stat.uiowa.edu/~rlenth/Power).

Animals were bred by and purchased from Charles River Germany (Sulzfeld, Germany). All mice were 7 weeks old upon arrival (body weight (g), mean ± standard error of the mean (SEM) and range: C, 20.4 ± 0.20 and 13.5 - 23.4; B6N, 21.0 ± 0.20 and 17.5 - 24.0; 129S2, 24.1 ± 0.27 and 19.3 - 28.3). Animals were

housed at the Central Laboratory Animal Research Facility of Utrecht University. Testing took place in the same rooms as where the animals were housed, and test equipment was placed in each room prior to arrival of the animals.

Mice were housed individually to reduce aggression and to avoid a potential confounding effect of aggression in (part of) the study population [45][46]. Mice were housed in Macrolon Type II L cages (size: 365 x 207 x 140 mm, floor area 530 cm^2, Techniplast, Milan, Italy) with standard bedding material (autoclaved Aspen Chips, Abedd-Dominik Mayr KEG, Köflach, Austria) and a tissue (KLEENEX® Facial Tissue, Kimberley-Clark Professional BV, Ede, the Netherlands) and a plastic PVC shelter as enrichment (Plexx BV, Elst, the Netherlands). Food (CRM, Expanded, Special Diets Services Witham, UK) and tap water were available *ad libitum*. Upon arrival mice were randomly allocated to one of two laboratory animal housing rooms for a habituation period of 17 days under a reversed 12 h light/12 h dark cycle (lights off at 7:00 AM) with a radio playing constantly as background noise. The number of mice per strain was similar between the two testing rooms. Relative humidity (mean percentage ± SEM) was controlled (room A: 53.5% ± 2.44; room B: 54.8% ± 2.56) with a ventilation rate of 15-20 changes/hour (both rooms) an average room temperature (mean ^0C ± SEM) of 21.7 ^0C ± 0.23 and 21.9 ^0C ± 0.40 for room A and B, respectively. The mice were handled three times a week during the habituation period by the same two experimenters that conducted the behavioral observations. Handling mice included picking up the mouse at the base of the tail and placed briefly on top of the home cage or on the arm of the experimenter to accustom them to test procedures.

2.3. Modified Hole Board

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

Figure 1 presents a schematic overview of the mHB. The apparatus consists of a grey PVC opaque box (100 x 50 x 50 cm) with a board made of the same material (60 x 20 x 20 cm) functioning as an unprotected area, as it is positioned in the center of box. The board stacks 20 cylinders (diameter 15 mm) in three lines.

The area around the board is divided into 10 rectangles (20 x 15 cm) and 2 squares (20 x 20 cm). The periphery was illuminated with red light (1-5 lux) and functioned as the protected area. In contrast, the central board was illuminated by an additional stage light in order to increase the aversive nature of the central (unprotected) area. Light intensity (mean lux) was 147 and 151 lux in room A and B, respectively.

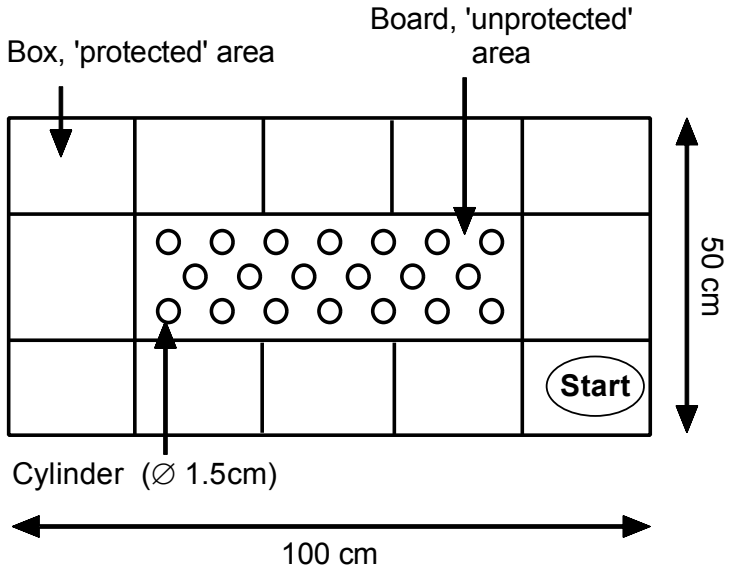


Figure 1. Schematic overview of the modified Hole Board.

2.4. Experimental protocol

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

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

At the start of the first trial, mice were transferred from the home cage to the mHB and always placed in the same corner, facing the central board. During the test, mice were allowed to freely explore the mHB-set up. Between trials mice were picked up by the tail and transported back to their home cage. The mHB was subsequently carefully cleaned with water and a damp towel before the next trial commenced. Behavior was scored live using the software Observer version 12.5 (Noldus Technology, Wageningen, the Netherlands). In addition, trials were recorded on video camera for raw data storage. Behavioral observations were conducted by two trained observers, each of which always tested in the same housing room. For obvious reasons, it was not possible to perform the observations blinded with respect to strain (due to the coat color of the animals, see section 1.2. Animals and Housing). Inter-observer reliability was established at a strong level [49] with an average Cohen's κ = 0.80 (range 0.71-0.95) over a percentage agreement of 84.27% (range 76.68 – 96.35) for frequency scores. For duration scores the inter-observer reliability was established at a strong level, with an average Cohen's κ = 0.88 (range 0.79-0.95) over a percentage agreement of 92.45% (range 86.20-97.28).

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

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

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

Within each pair, one mouse was treated with an anxiolytic, while the other served as control. Treatment and control were assigned randomly within pairs. Treatment consisted of an intra-peritoneal injection (*i.p.*) with dexmedetomidine (Dexdomitor®, 10 µg/kg, 100 µl; Orion Corporation-Orion Pharma, Espoo, Finland). Often used as a sedative-analgesic agent, this pharmacon is a highly selective α_2-adrenergic receptor antagonist that also poses anxiolytic properties [50]. The selected dose was based on a pilot study in which this dose produced behavioral changes but no sedative effect. Control mice of each pair received a saline injection (NaCl, 0.9%, 100 µl, *i.p*). Treatment and control were assigned randomly within pairs. Pairs of mice were tested over a period of 4 consecutive days, with a weekend in between (thu-fri-mo-tue). The experiment was designed as a complete randomized block design in which each test day was treated as a separate block. The numbers of pairs were maintained equal between test days, between strains and between experimenter. This amounted to 1 balanced, and 1 unbalanced pair (2) per strain (3), per experimenter (2) per block (4), resulting in 48 pairs. Figure 2 presents a schematic representation of the distribution of pairs within a single block (test day), for one experimenter.

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

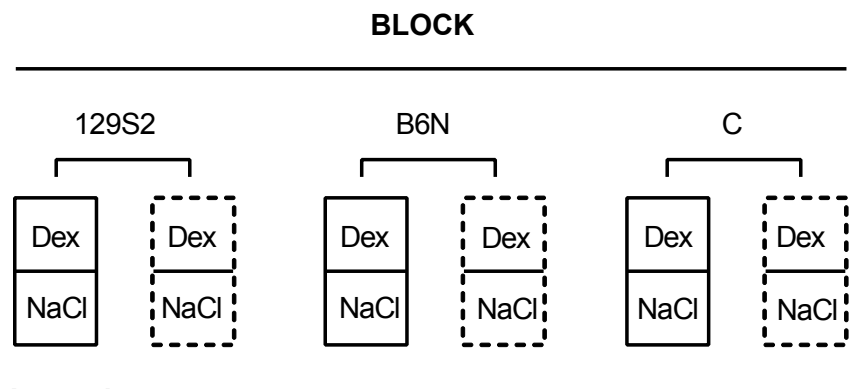


Figure 2. Schematic overview of the 2 (treatment) x 3 (strain) x 2 (experimenter) factorial complete randomized block design used in the pharmacological experiment. Overview represents a single block, for one experimenter. Pairs in the balanced condition were matched on weight and individual response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

2.5. Behavioral variables

Experimental phase 1 and 2: Behavioral patterns of mice were assessed by scoring behaviors listed in Table 1. From these observations, several parameters for avoidance behavior, exploration and locomotion were computed (Table 1). Furthermore, previous studies using the mHB have shown that these separate behaviors scored in this assay can be reliably summarized to underlying

behavioral dimensions; avoidance behavior, exploration and locomotion [51][52][53]. Two previous studies showed that simultaneous clustering of these behavioral dimensions yields distinct response types that are displayed by individuals of C, B6N and 129S2 [31][32]. For each dimension, observed behaviors indicative of that dimension were summarized to integrated behavioral z-scores according to the procedure described in Labots et al. [53] and van der Goot et al. [31]. In short, this entailed that behavioral variables measuring different aspects of the same behavioral dimension were normalized to z-scores, and combined to a single integrated z-score representing that dimension. Combination of z-scores was done by averaging them. For normalization of each separate variable, we used the pooled data (across all strains) as a reference group, as suggested by Labots et al. [53]. An overview of all included variables per dimension is listed in supplementary Table S1.

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

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

2.6. Statistics

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

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

The procedure described in section 4.5 yielded three trajectories of integrated residual z-scores for each individual mouse, one trajectory per behavioral dimension: avoidance behavior, exploration and locomotion. These three trajectories were fit with LMMs to control for potentially confounding factors. The resulting standardized Pearson residuals could then be used for a clustering procedure. For each behavioral dimension, potentially confounding variables were controlled for by including strain and experimenter as fixed factors, and individual mouse, test group and test order as random factors in the model. The variable 'trial' was intentionally left out of the model because we wanted to maintain this information in the residuals so that we could assess habituation of individual mice over time. Models were run with an autoregressive correlation structure for continuous time covariates (corCAR1) from the 'nlme' package.

Model assumptions were assessed visually by inspecting the standardized residuals through QQ-plots, histograms and residual plots [57][58]. The dimension avoidance behavior was logarithmically transformed to achieve normality of the residuals. A square root transformation was applied on exploration and locomotion was rank transformed. Heteroscedasticity was avoided using the 'varIdent' variance structure function from the 'nlme' package, allowing different residual spread for each level of the categorical variables in our model [58]. The dimensions avoidance behavior, exploration and locomotion included a variance function for 'Strain'.

The resulting standardized Pearson residual integrated z-score trajectories were subsequently analyzed with a multivariate k-means clustering procedure for longitudinal data, *kml3d* [56]. The settings and rationale for using this method have been described in detail in [31]. Furthermore, the settings used in the present manuscript are identical to a previous study [32]. As described, three response trajectories were included for each individual mouse: avoidance behavior, exploration and locomotion. These were clustered simultaneously to assess the occurrence of homogenous subgroups of mice that shared similar responses (between them) on all three behavioral dimensions. Prior to analysis, the gap statistic was applied to evaluate whether the data was perhaps best represented by a single cluster using the package 'cluster' [59]. This was not the case. The gap statistic compares the within-cluster sum-of-squares to a null reference distribution of the data, which is then equivalent to a single cluster [60], and as such gives an indication of whether it is appropriate to partition the data into clusters. The cluster analysis compiled 1000 iterations for each *k* clusters between 2 and 6, resulting in 5000 cluster solutions.

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

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

Main and interaction effects from all LMMs were derived using *F*-tests with corresponding *P* value ($P < 0.05$). Statistical significance of random effects was computed by means of likelihood ratio tests and reported as Chi Square values. Pairwise comparisons were conducted using the package 'emmeans' [64] to follow up on main or interaction effects. To reduce the probability of a Type I error due to multiple comparisons, the α was adjusted using a Dunn-Šidák correction in all *post hoc* tests [65]. Supplementary Table S2 lists an overview of all corrected α-values used in this manuscript. All *post hoc* tests were summarized as beta-estimates and their corresponding standard error, *t* statistic and *P* values. Effect sizes for *post hoc* tests were reported as Cohen's *d*, and obtained via the package 'emmeans' [64]. The guidelines provided by Wahlsten [66] were used to interpret the absolute values of Cohen's *d* (|d|). This extensive review of various phenotypes suggested the following interpretation of effects for neurobehavioral mouse studies: small effect, $|d| < 0.5$; medium effect, $0.5 < |d| < 1.0$; large effect, $1.0 < |d| < 1.5$; very large effect, $|d| > 1.5$.

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

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

For all analyses in phase 2, the integrated behavioral z-score for exploration was logarithmically transformed and that of locomotion was rank transformed to improve the residual distribution. In addition, Cooks distance identified locomotion scores of 4 mice as influential – these were 2 individuals that had not displayed any line crossing, resulting in the maximum score for latency to first line crossing (300 seconds) and 2 individuals with a latency to first line crossing > 180 seconds. These observations were retained for analysis. Main and interaction effects from all GLMs were derived using *F*-tests with corresponding *P* value ($P < 0.05$). Effect sizes for the GLMs are reported as partial eta squared (η_p^2) with 95% CI, using the following cut-off limits: small effect, $\eta_p^2 \leq 0.03$; medium/moderate effect, $0.03 < \eta_p^2 < 0.10$; large effect, $0.10 \leq \eta_p^2 < 0.20$; very large effect, $\eta_p^2 \geq 0.20$ [68]. Pairwise comparisons were conducted using the package 'emmeans' [64] to follow up on main or interaction effects. To reduce the probability of a Type I error due to multiple comparisons, the α was adjusted using a Dunn-Šidák correction in all *post hoc* tests [65], see supplementary Table S2. All *post hoc* tests were summarized as beta-estimates and their corresponding standard error, *z* statistic and *P* values. Effect sizes for *post hoc* tests were reported as Cohen's *d*, and obtained via the package 'emmeans' [64]. The guidelines provided by Wahlsten [66] were again used to interpret the absolute values of Cohen's *d* (|d|), see *"Experimental Phase 1"*.

3. Results

3.1. Cluster analysis

The optimal partitioning of the data yielded two clusters, A and B. The table in Figure 3a presents cluster size and distribution of strains across clusters. The majority of individuals (57.5%, n = 103) grouped together in cluster A while the remaining mice formed cluster B (42.5%, n = 76). Each cluster consisted of mice from all three strains. The majority of C mice (88.1%, n = 52) grouped together in cluster B while the majority of 129S2 (90%, n = 54) and the majority of B6N (70%, n = 42) grouped together in cluster A. The clusters displayed differential patterns of behavior on all three behavioral dimensions, as indicated by significant interactions between clusters and trial (Avoidance behavior, trial effect: $F_{(4,708)} = 13.17$, $P < 0.0001$; interaction cluster x trial: $F_{(4,708)} = 46.58$, $P < 0.0001$; Exploration, cluster effect: $F_{(1,177)} = 12.12$, $P = 0.0006$; trial effect: $F_{(4,708)} = 37.63$, $P < 0.0001$; interaction cluster x trial: $F_{(4,708)} = 44.97$, $P < 0.0001$; Locomotion, cluster effect: $F_{(1,177)} = 78.43$, $P < 0.0001$; trial effect: $F_{(4,708)} = 4.85$, $P = 0.0007$; interaction cluster x trial: $F_{(4,708)} = 19.64$, $P < 0.0001$; Figure 3b).

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

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

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

a.

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

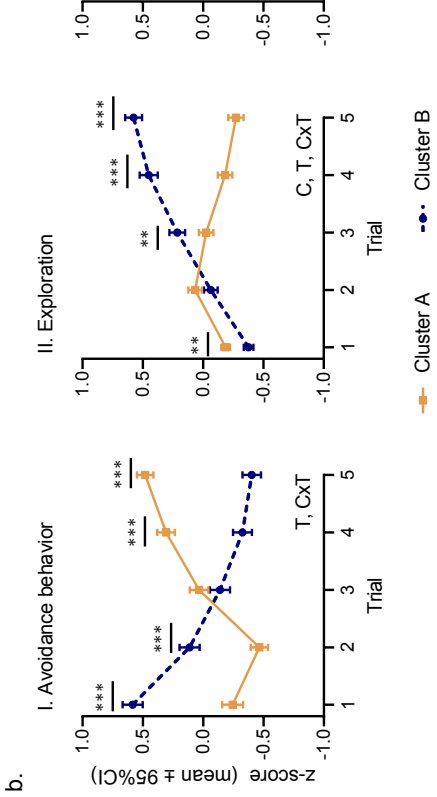


Figure 3. (a) Results cluster analysis. Top row: Cluster size and proportion of total population for each cluster (total number of mice, *n* = 179). Bottom rows: Distribution of strains (*n* and proportion) within each cluster. **(b)** Behavioral response trajectories of clusters on avoidance behavior,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

supplementary Table S4-II). Locomotion was significantly higher in cluster B on all trials except trial 1 (trial 2, $P < 0.0001$, *medium* effect size, d = -0.774, 95%CI [-1.140, -0.406]; trial 3, $P < 0.0001$, *large* effect size, d = -1.157, 95%CI [-1.510, -0.798]; trial 4, $P < 0.0001$, *large* effect size, d = -1.384, 95%CI [-1.730, -1.033]; trial 5, $P < 0.0001$, *very large* effect size, d = -1.562, 95%CI [-1.910, -1.208], Figure 3b-III, supplementary Table S4-II).

3.1.1 Cluster stability

Cluster stability was assessed with a bootstrapping procedure in which 200 random samples (of n = 179) were drawn from the original data with replacement. The trajectories of the original cluster and the average of all 200 cluster analyses were highly similar (Figure 4a). The average Jaccard Index was 0.92 for cluster A (Figure 4b), meaning that on average, mice that fell in cluster A also did so in 92% of the bootstrap samples. The average Jaccard Index for cluster B was equally high (0.95, Figure 4b). The originally obtained clusters A and B were thus rendered stable, and representative for this dataset.

3.2. Results pharmacological experiment

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

3.2.1 Incorporating individual variation as a discriminating factor in analysis

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

Treatment with dexmedetomidine primarily reduced activity related behavior, compared to a control injection with saline. Treated mice displayed less exploration ($F_{(1,68)}$ = 7.36, P = 0.0085, *medium* effect size, η_p^2 = 0.097, 95% CI [0.008, 0.252]) and less locomotion than controls ($F_{(1,68)}$ = 9.75, P = 0.0027, *large*

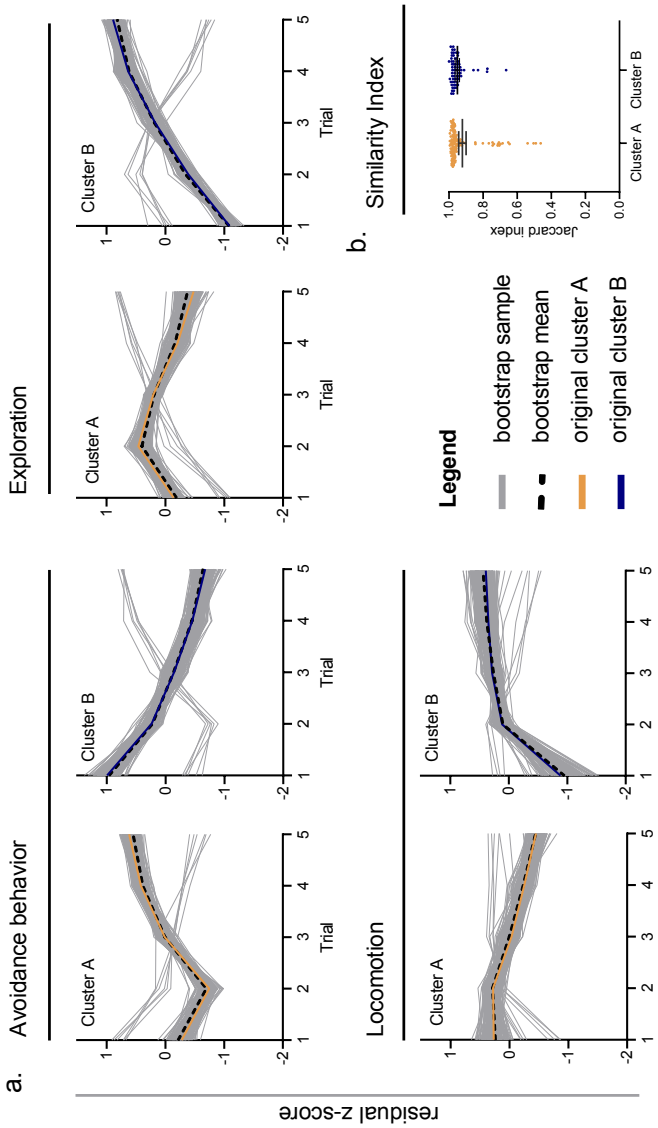


Figure 4. (a) Results of the bootstrapping procedure for each cluster, on each behavioral dimension. Results are presented as mean residual z-scores, and depict the trajectory of the original cluster (cluster A, orange; cluster B, blue) in relation to the average of all 200 bootstrap samples (black dashed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

effect size, η_p^2 = 0.124, 95% CI [0.016, 0.279]; Figure 5a-II, 5a-III), supplementary Table S5). The effect of dexmedetomidine on anxiety related behavior was less pronounced, but there was a suggestion of average higher levels of avoidance behavior in treated animals ($F_{(1,68)}$ = 3.70, *P* = 0.0586, *medium* effect size, η_p^2 = 0.052, 95% CI [0.000, 0.185]; Figure 5a-I, supplementary Table S5).

Strain also differed significantly in exploration ($F_{(2,68)}$ = 13.38, *P* = 0.0000) and locomotion ($F_{(2,68)}$ = 29.23, *P* < 0.0001), both with very large effect sizes (respectively η_p^2 = 0.282, 95%CI [0.104, 0.423]; η_p^2 = 0.493, 95%CI [0.319, 0.616]; Figure 5b-II, 5b-III, supplementary Table S5). *Post hoc* comparisons (adjusted α = 0.02532) showed that exploration was highest in B6N compared to C and 129S2 (C, *P* = 0.0002, *medium* effect size, *d* = -0.928, 95%CI [-1.499, -0.408]; 129S2, *P* < 0.0001, *large* effect size, *d* = -1.327, 95%CI [-1.891, -0.763]; Figure 5b-II, supplementary Table S6a). Furthermore, locomotion differed between all three strains, with higher levels of locomotion in B6N compared to C and 129S2 (C, *P* < 0.0001, *large* effect size, *d* = -1.240, 95%CI [-1.780, -0.700]; 129S2, *P* < 0.0001, *very large* effect size, *d* = -2.000, 95%CI [-2.620, -1.381]) and higher locomotion in C than in 129S2 (*P* = 0.0028, *medium* effect size, *d* = -0.760, 95%CI [-1.270, -0.246], Figure 5b-III, supplementary Table S6a).

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

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

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

3.2.2 Comparison between balanced and unbalanced pool

Next, we analyzed the balanced and the unbalanced pool separately and compared the results. Aside for controlling for individual response type, we considered the two pools directly comparable with respect to other factors such as experimenter, strain, treatment etcetera. Any difference in results between these two pools of mice was therefore attributed to the fact that we matched our pairs on individual response type in one pool (balanced) and not in the other (unbalanced). Furthermore, following the principle of good experimental design described in the introduction, matching our pairs on individual response types in the balanced pool should make the treatment and control groups more similar, thereby improving the quality of our results. Any differences in results between the balanced and the unbalanced pool were thus interpreted in favor of the balanced data.

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

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

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

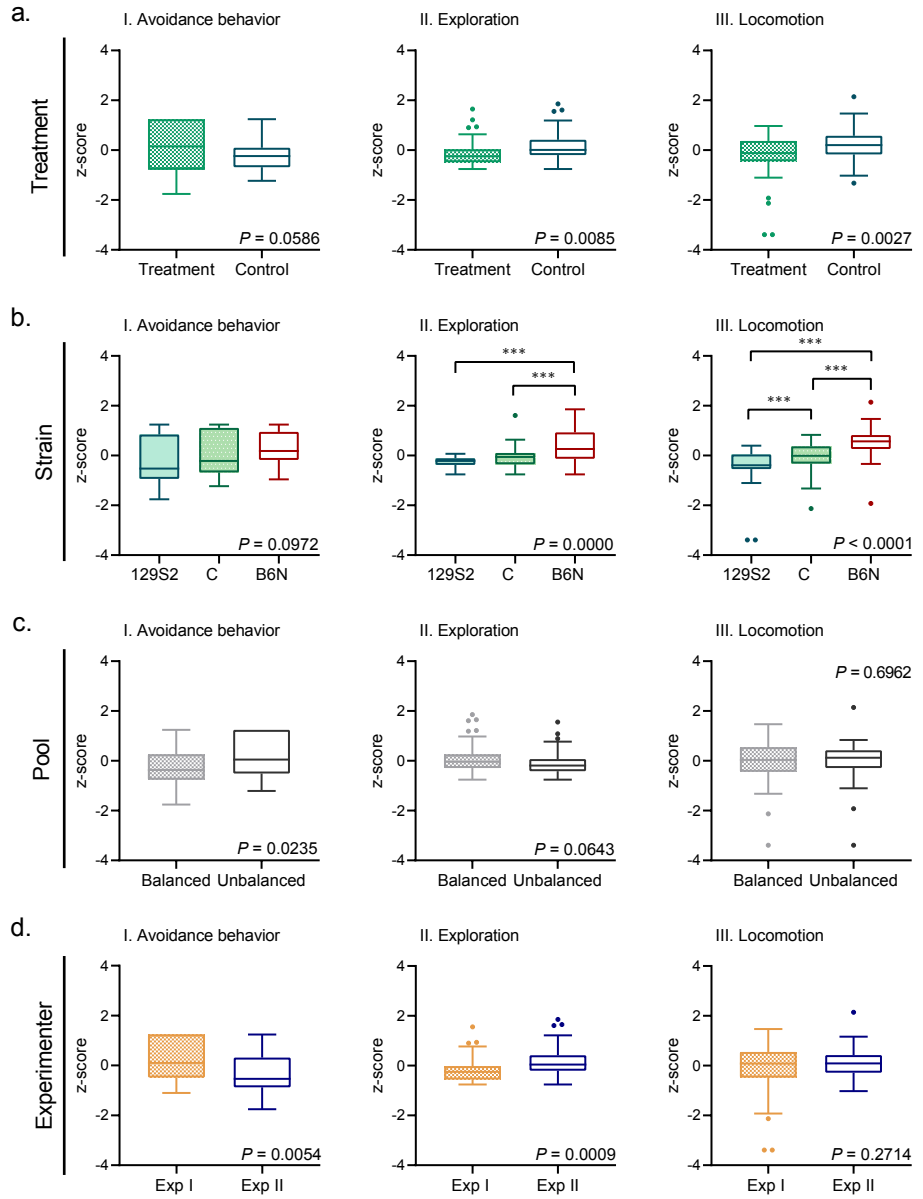


Figure 5. (a-d) Avoidance behavior, exploration and locomotion of mice in a single 5-minute mHB trial. Behavior in all graphs expressed as integrated behavioral z-score. Results are presented as boxplots (median, upper and lower quartiles) with Tukey whiskers. Individual points = outside the upper/lower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

[-3.691, -1.371]; Figure 6II-1, supplementary Table S6c). Furthermore, *post hoc* comparisons between conditions within strain (adjusted α = 0.016952) showed that saline injected B6N displayed more exploration than their counterparts treated with dexmedetomidine (*P* < 0.0001, *very large* effect size, *d* = 2.105, 95%CI [0.997, 3.213]; Figure 6II-1, supplementary Table S6c).

Interestingly, this treatment effect disappeared when analyzing the balanced pool. Exploration still differed between strains (strain effect: $F_{(2,32)}$ = 8.24, *P* = 0.0013, *very large* effect size, η_p^2 = 0.340, 95%CI [0.070, 0.518]) but exploration was now higher in B6N than C and 129S2 regardless of treatment, as opposed to only in the saline condition (C, *P* = 0.0031; *large* effect size, *d* = -1.053, 95%CI [-1.800, -0.310]; 129S2, *P* = 0.0001, *very large* effect size, *d* = -1.719, 95%CI [-2.690, -0.743]; Figure 6II-2, supplementary Table S6b). Also, the treatment effect within B6N disappeared in the balanced condition (Figure 6II-2).

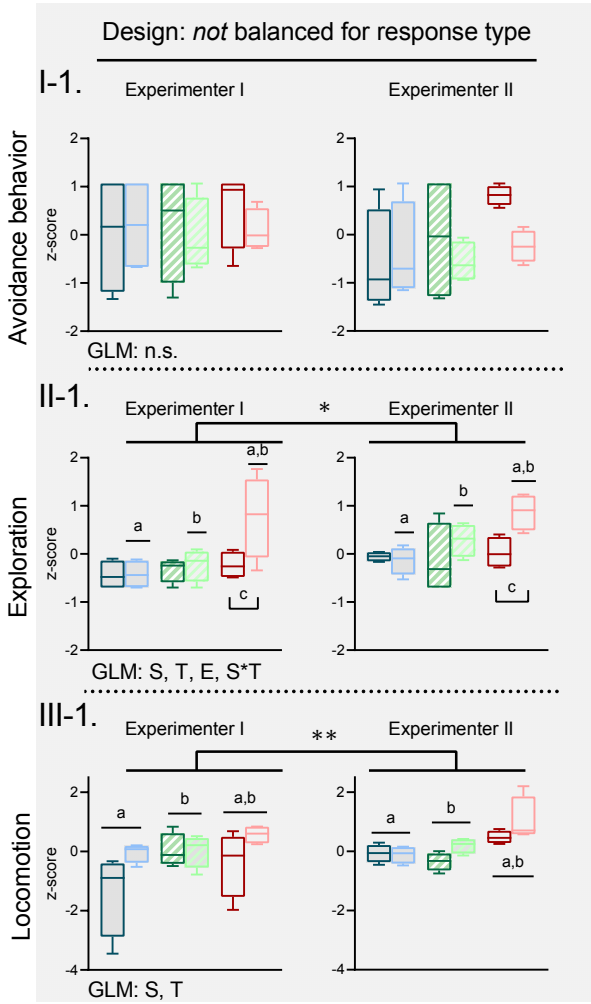
A similar shift of the effect of treatment was found for locomotion. In the unbalanced pool, locomotion was significantly higher in controls compared to treated mice across strains ($F_{(1,32)}$ = 10.15, *P* = 0.0032, *very large* effect size, η_p^2 = 0.241, 95%CI [0.032, 0.450]; Figure 6III-1, supplementary Table S7) but when controlling for individual response type this treatment effect disappeared in the balanced pool ($F_{(1,32)}$ = 1.62, *P* = 0.2224, *medium* effect size, η_p^2 = 0.046, 95%CI [0.000, 0.249]; Figure 6III-2, supplementary Table S7). Similar to exploration, strains differed in locomotion in both the balanced ($F_{(2,32)}$ = 13.78, *P* = 0.0002, *very large* effect size, η_p^2 = 0.463, 95%CI [0.176, 0.616]; Figure 6c-II) and the unbalanced pool ($F_{(2,32)}$ = 12.05, *P* = 0.0001, *very large* effect size, η_p^2 = 0.430, 95%CI [0.144, 0.590]; Figure 6III-1). *Post hoc* comparisons (adjusted α = 0.02532) showed that locomotion (rank transformed) was significantly higher in B6N than in C and 129S2 in both the unbalanced (C, *P* = 0.0021, *large* effect size, *d* = -1.116, 95%CI [-1.880, -0.354]; 129S2, *P* < 0.0001, *very large* effect size, *d* = -1.713, 95%CI [-2.520, -0.903]; Figure 6c-I, supplementary Table S6c) and balanced condition (C, *P* = 0.0002, *large* effect size, *d* = -1.320, 95%CI [-2.090, -0.520]; 129S2, *P* < 0.0001, *very large* effect size, *d* = -2.250-, 95%CI [-3.290, -1.209], Figure 6III-2, supplementary Table S6b).

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

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

4. Discussion

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



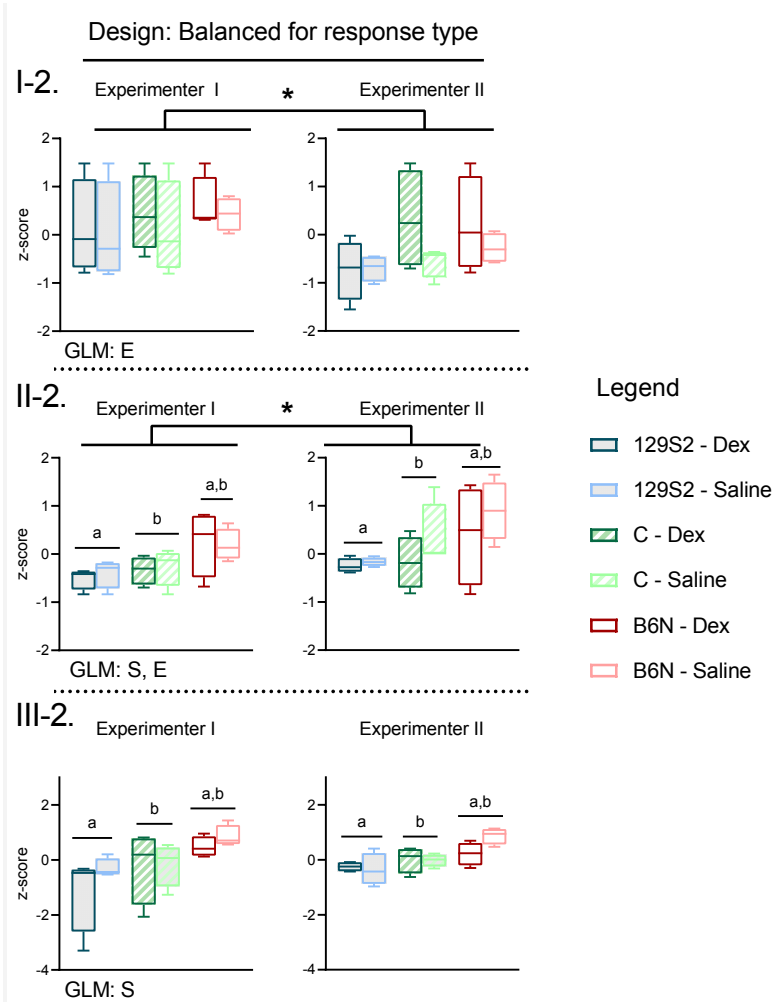


Figure 6. Behavioral scores of a pool of mice not balanced for individual response type (left), and a pool balanced for individual response type (right). **(I-III)** Results expressed as integrated behavioral z-scores and presented as boxplots (median, upper and lower quartiles) with Tukey whiskers.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

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

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

At the same time however, we recognize that the procedure of a priori characterization may not be suitable or desirable for all phenotypes or contexts. In behavioral neurosciences for example, initial testing for the purpose of individual characterization may be undesirable because the animals need to be naïve to the test, although evidence suggests that naivety to the test can be

ensured by allowing sufficient time between characterization and test [68][69]. Second, characterizing individuals and subsequently using this characterization in experimental design requires that the observed trait is consistent across time. With respect to anxiety, individuality in both anxiety-related and activity behaviors have indeed often been shown to be repeatable and consistent through time and context in isogenic mice [27][28][29][70][71][72]. Other behavioral traits however, such as grooming behavior, have been shown to be less consistent across time within the same individual [72]. In the present study we unfortunately did not test the temporal consistency of our individual response types. The identified clusters were stable however at the time of assessment (Fig. 2) and the behavioral profiles of the response types largely overlapped with two previous studies [31][32], suggesting some consistency. This aspect requires further validation, as confirmation of temporal stability of our response types will substantiate our claim that the difference in results between a balanced and the unbalanced pool can be attributed to variation related to inter-individual differences in response.

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

A highly efficient alternative to a priori characterization is the use of cross-over designs [75][76]. In these designs, for example a balanced Latin Square design, variability between individuals is accounted for by contrasting all treatments within the same animal [76]. Each animal thereby serves as its own control, which reduces the sample size while maintaining the same statistical power [76]. With respect to the current study, it should be noted that this would indeed have been a more practical and time efficient strategy to account for inter-individual variability if evaluating the effectiveness of dexmedetomidine as an anxiolytic would have been the sole purpose of this paper. As described

above however, our primary aim was not so much to study dexmedetomidine as an anxiolytic, but to evaluate whether balancing experimental groups with respect to inter-individual variability would yield different results compared to a 'regular' drug study in which this variability is not actively accounted for.

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

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

Next, in addition to these methodological and statistical considerations, this study confirmed previously identified inter-individual differences in the ability to habituate anxiety-related behavior in C, 129S2 and B6N [32]. The profiles were characterized by differential patterns of avoidance behavior and

exploration: Mice in Cluster A combined an increase of avoidance behavior with low levels of exploration that remained stable over trials, while avoidance behavior decreased and exploration increased in Cluster B (Fig. 3). This interplay between anxiety and exploration may be explained by the so-called approach-avoidance conflict that exposure to unprotected areas may induce in rodents, which entails the motivational conflict between the drive to explore a novel environment and the motivation to avoid potentially harmful stimuli [13][33][34]. Following this interplay, the decrease in avoidance behavior in Cluster B may be interpreted as successful habituation of anxiety responses, while the increase in avoidance in Cluster A may be considered as impaired habituation to the test.

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

Furthermore, this study presented marked experimenter differences in behavioral scores for avoidance behavior and exploration, despite the fact that the experiment was carefully balanced with respect to experimenter. Behavioral phenotyping of anxiety-related behavior in rodents has indeed been demonstrated to be sensitive to experimenter-related factors such as handling style and familiarity with the experimenter [84][85]. In this study, both experimenters were naïve to handling rodents and were trained by the same person to handle the mice. Furthermore, all animals were handled by both experimenters from arrival at the test facility onwards. These measures however unfortunately do not preclude the possibility that handling differences or other experimenter-related factors affected the outcome of this study as experimenters themselves can never be entirely subjected to standardization [86]. Another experimenter-induced factor that may affect the scoring of live behavior is observer variability [38][87][88]. Both inter- and intra-observer

reliability were established at a good to excellent level prior to the start of the study, after a training phase in which both experimenters aligned coding by scoring video data from previously collected mHB-data. Observer reliability in itself may however again be affected by coding experience, rapidity of the behavior, energy level of the observer and so on [11][89]. These factors illustrate how complete control over experimenter-induced variability is difficult to accomplish. In fact, the experimenter has been termed one of the most uncontrollable background factors in experimental research, affecting experimental outcomes and reproducibility between studies in a similar manner as inter-individual variability [38].

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

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

Conclusions

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

Acknowledgements

The authors would like to dedicate this article to the memory of prof. dr. Frauke Ohl. We remember Frauke as a beloved colleague and supervisor and we are most grateful for inspiring us to continue the work leading up to this manuscript. In addition, the authors are grateful to Dr. J.R. Yates and an anonymous reviewer for their valuable comments on the manuscript, which has undoubtedly helped the authors to improve the article.

References

1. Koolhaas JM, de Boer SF, Coppens CM, Buwalda B. Neuroendocrinology of coping styles: Towards understanding the biology of individual variation. Front. Neuroendocrinol. 2010; 31, 307-321.
2. Gärtner K. A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 yearlong effort to standardize laboratory animals? Int. J. Epidemiol. 2012; 41, 335-341. Reprint of Lab. Anim. 1990; 24**,** 71-77.
3. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in the light of biological variation. Nat. Rev. Neurosci. 2020; 21, 384-393.
4. Lathe R. The individuality of mice. Genes Brain Behav. 2004; 3, 317–327.
5. Einat H, Ezer I, Kara N, Belzung C. Individual responses of rodents in modelling of affective disorders and in their treatment: prospective review. Acta Neuropsychiatr. 2018; 30, 323-333.
6. Lewejohann L, Zipser B, Sachser N. „Personality" in laboratory mice used for biomedical research: A way of understanding variability? Dev. Psychobiol. 2001; 53 (6), 624-630.
7. Richter SH, Garner JP, Auer C, Kunert J, Wuerbel H. Systematic variation improves reproducibility of animal experiments. Nat. Meth. 2010; 7, 167-168.
8. Richter SH. Systematic heterogenization for better reproducibility in animal experimentation. Lab. Anim. 2017; 46, 343-349.
9. Voelkl B, Würbel H. A reaction norm perspective on reproducibility. Preprint at http:/bioRxiv.org/content/10.1101/510941v3, 2020.
10. Kafkafi N, Agassi J, Chesler EJ. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci. Biobehav. Rev. 2018; 87, 218-232.
11. Bello NM, Renter DG. Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. J. Dairy Sci. 2018; 101, 5679-5701.
12. Rojas-Carvajal M, Quesada-Yamasaki D, Brenes JC. The cage test as an easy way to screen and evaluate spontaneous activity in preclinical neuroscience studies. Methodsx 2021; 8, 101271.
13. Armario A, Nadal R. Individual differences and the characterization of animal models of psychopathology: a strong challenge and a good opportunity. Front. Pharmacol. 2013; 4, 137.
14. Lonsdorf TB, Merz CJ. More than just noise: Inter-individual differences in fear acquisition, extinction and fear in humans – Biological, experiential, temperamental factors, and methodological pitfalls. Neurosci. Biobehav. Rev. 2017; 80, 703-728.
15. Cohen H, Geva AB, Matar MA, Zohar J, Kaplan Z. Post-traumatic stress behavioural responses in inbred mouse strains: can genetic predisposition explain phenotypic variability? Int. J. Neuropsychoph. 2008; 11, 331-349.
16. Galatzer-Levy IR, Bonanno GA, Bush DEA, LeDoux JE. Heterogeneity in threat extinction learning: substantive and methodological considerations for identifying individual differences in response to stress. Front. Behav. Neurosci. 2013; 7, 55.
17. Pawlak CR, Ho Y, Schwarting RKW. Animal models of human psychopathology based on indivdiual differences in novelty-seeking and anxiety. Neurosci. Biobehav. Rev. 2008; 32, 1544-1568.
18. Harro J. Inter-individual differences in neurobiology as vulnerability factors for affective disorders: Implications for psychopharmacology. Pharmacol. Ter. 2010; 125 (3), 402-422.
19. Barbelivien A, Billy E, Lazarus C, Kelche C, Majchrzak M. Rats with different profiles of impulsive choice behavior exhibit differences in responses to caffeine and d-amphetamine and in medial prefrontal cortex 5-HT utilization. Behav. Brain. Res. 2008; 187 (2), 273-282.
20. Festing MFW. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. ILAR Journal 2014; 55, 399-404.
21. Festing MFW. Study Design. In: Martin-Kehl, MI, Schubiger, PA, editors. Animal Models for Human Cancer: Discovery and Development of Novel Therapeutics. Wiley-VCH, Weinheim; 2016, pp. 27-40.
22. Guiliano C, Peña-Liver Y, Goodlett CR, Cardinal RN, Robbins TW, Bullmore ET et al. Evidence for a long-lastig compulsive alcohol seeking phenotype in rats. Neuropsychopharmacology 2018; 43, 728-738.
23. Giuliano C, Puaud M, Cardinal RN, Belin D, Everitt BJ. Individual differences in the engagement of habituation control over alcohol seeking predicts the development of compulsive alcohol seeking and drinking. Preprint at http:/bioRxiv.org/10.1111/adb13041, 2021.
24. Irwin JR, McClelland GH. Negative consequences of dichotomizing continuous predictor variables. J. Mark. Res. 2003; 40 (3), 366-371.
25. Stegman Y, Schiele MA, Schümann D, Lonsdorf TB, Zwanzger P, Romanos M, et al. Individual differences in human fear generalization – pattern identification and implications for anxiety disorders. Transl. Psychiatry 2019; 9, 307.
26. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 2010; 167 (7), 748–751.
27. Freund J, Brandmaier AM, Lewejohann L, Kirste I, Kritzler M, Krüger A, et al. Emergence of individuality in genetically identical mice. Science 2013; 340, 756-759.
28. Keshavarz M, Krebs-Wheaton R, Refki P, Savriama Y, Zhang Y, Guenther A, et al. Natural copy number variation differences of tandemly repeated small nucleolar RNAs in the Prader-Willi syndrome genomic region regulate individual behavioral responses in mammals. Preprint at http:/bioRxiv.org/content/10.1101/476010v2, 2020.
29. Kazavchinsky L, Dafna A, Einat H. Individual variability in female and male mice in a test-retest protocol of the forced swim test. J. Pharmacol. Toxicol. Methods 2019; 95**,** 12-15.
30. Tuttle AH, Philip VM, Chesler EJ, Mogil JS. Comparing phenotypic variation between inbred and outbred mice. Nat. Methods 2018; 15 (12), 994-996.
31. van der Goot MH, Boleij H, van den Broek J, Salomons AR, Arndt SS, van Lith HA. An individual based, multidimensional approach to identify emotional reactivity profiles in inbred mice. J. Neurosci. Meth. 2020; 343, 108810.

32. van der Goot MH, Keijsper M, Baars A, Drost L, Hendriks J, Kirchhoff S, Lozeman-van t Klooster JG, van Lith HA, Arndt SS. Inter-individual variability in habituation of anxiety-related responses within three mouse inbred strains. Phys. Behav. 2021; 113503 (Epub ahead of print).
33. O'Leary TP, Gunn RK, Brown RE. What are we measuring when we test strain differences in anxiety in mice? Behav. Genet. 2013; 43 (1), 34–50.
34. Ohl F. Testing for anxiety. Clin. Neurosci. Res. 2003; 3 (4-5), 233-238.
35. Belzung C, Griebel G. Measuring normal and pathological anxiety-like behavior in mice: a review. Behav. Brain Res. 2001; 125 (1-2), 141-149.
36. Festing MFW. Experimental design and irreproducibility of pre-clinical research. Physiol. News 2020; 118, 14-15.
37. Festing MFW. The "completely randomized" and the "randomized block" are the only experimental designs suitable for widespread use in pre-clinical research. Sci. Rep. 2020; 10, 17577.
38. Richter SH. Automated home cage testing as a tool to improve reproducibility of behavioral research? Front. Neurosci. 2020; 14, 383.
39. Salomons AR, Espitia Pinzon N, Boleij H, Kirchhoff S, Arndt SS, Nordquist RE, et al. Differential effects of diazepam and MPEP on habituation and neurobehavioral processes in inbred mice. Behav. Brain Funct. 2012; 8, 30.
40. Weerink MAS, Struys MMRF, Hannivoort LN, Barends CRM, Absalom AR, Colin P. Clinical pharmacokinetics and pharmacodynamics of dexmedetomidine. Clin. Pharmacokinet. 2017; 56, 893-913.
41. Laarakker MC, van Raai JR, van Lith HA, Ohl F. The role of the alpha 2A-adrenoceptor in mouse stress-coping behaviour. Psychoneuroendicronology 2010; 35, 490-502.
42. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. BMC Vet Res. 2020; 16, 242.
43. Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLos Biol. 2020; 18, e3000411.
44. Meziane H, Quagazzal A-M, Aubert L, Wietrzych M, Krezel W. Estrous cycle effects on behavior of C57BL/6J and BALB/cByJ female mice: implications for phenotyping strategies. Genes Brain Behav. 2007; 6, 192-200.
45. Arndt SS, Laarakker MC, van Lith HA, van der Staay FJ, Gieling E, Salomons AR, et al. Individual housing of mice – Impact on behavior and stress responses. Phys. Behav. 2009; 97 (3-4), 385-393.
46. Kappel S, Hawkins P, Mendl MT. To group or not to group? Good practice for housing male laboratory mice. Anim. 2017; 7 (12), 88.
47. Ohl F, Holsboer F, Landgraf R. The modified hole board as a differential screen for behavior in rodents. Behav. Res. Methods Instr. Comput. 2001; 33 (3), 392-397.

48. Labots M, van Lith HA, Ohl F, Arndt SS. The modified hole board –measuring behavior, cognition and social interaction in mice and rats. J. Vis. Exp. 2015; 98**,** e52529.
49. Cicchetti DV. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsyc. 2001; 23, 695-700.
50. Gertler R, Brown C, Mitchell D, Silvius E. Dexmedetomidine: a novel sedative-analgesic agent. Proc. (Bayl. Univ. Med. Cent.) 2001; 14, 13-21.
51. Laarakker MC, Ohl F, van Lith HA. Chromosomal assignment of quantitaive trait loci influencing modified hole board behavior in laboratory mice using consomic strains, with special reference to anxiety-related behavior and mouse chromosome 19. Behav. Genet. 2008; 38, 159-184.
52. Laarakker MC, van Lith HA, Ohl F. Behavioral characterization of A/J and C57BL/6J mice using a multidimensional test: association between blood plasma and brain magnesium-ion concentration with anxiety. Physiol. Behav. 2011; 102, 205-219.
53. Labots M, Laarakker MC, Schetters D, Arndt SS, van Lith HA. An improved procedure for integrated behavioral z-scoring illustrated with modified Hole Board behavior of male inbred laboratory mice. J. Neurosci. Methods 2018; 293, 375-388.
54. R Core Team. R: A language environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria; 2020.
55. Pinheiro J, Bates D, Debroy S, Sarkar D, R Core Team. nlme: Linear and nonlinear mixed effects models. R package version 3.1.-147; 2020.
56. Genolini C, Alacoque X, Sentenac M, Arnaud C. Kml and kml3d: R-packages to cluster longitudinal data. J. Stat. Softw. 2015; 65, 1-34.
57. Sokal RR, Rohlf FJ. Biometry: The Principles and Practice of Statistics in Biological Research. 3rd ed. New York, NY: W.H. Freeman and Co. 1995.
58. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York, NY: Springer; 2005.
59. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0; 2019.
60. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B. 2002; 63, 411-423.
61. Kryszczuk K, Hurley P. Estimation of the number of clusters using multiple clustering validity indices. In: El Gayar N, Kittler J, Roli F, editors. Multiple Classifier Systems. Lecture notes on Computer Science, vol 5997, New York, NY: Springer; 2010, pp 114-123.
62. Wahl S, Krug S, Then C, Kirchhofer A, Kastenmüller G, Brand T, et al. Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele. Metabolomics 2014; 10, 386-401.
63. Clatworthy J, Buick D, Hankins M, Weinman J, Home R. The use and reporting of cluster analysis in health psychology: a review. Br. J. Health. Psychol. 2005; 10, 329-358.

64. Lenth R. Emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.7; 2020.
65. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. 1967; 62, 626-633.
66. Wahlsten D. Sample size. In: Wahlsten D, Mouse Behavioral Testing: How to Use Mice in Behavioral Neuroscience, London, UK: Academic Press, Elsevier Inc.; 2011, pp 75-105.
67. Festing MFW. The principles of experimental design and the determination of sample size when using animal models of traumatic brain injury. In: Srivastava A, Cox C, editors. Pre-Clinical and Clinical Methods in Brain Trauma Research, vol 139; New York, NY: Humana Press, 2018; pp. 201-225.
68. Labots M, Zheng X, Moattari G, Ohl F, van Lith HA. Effects of light regime and substrain on behavioral profiles of male C57BL/6 mice in three tests of unconditioned anxiety. J. Neurogenet. 2016; 30, 306-315.
69. Bouwknecht JA, Paylor R. Pitfalls in the interpretation of genetic and pharmacological effects on anxiety-like behaviour in rodents. Behav. Pharmacol. 2008; 19, 385-402.
70. Jakovcevski M, Schachner M, Morellini F. Individual variability in the stress response of C57BL/6J male mice correlates with trait anxiety. Genes Brain Behav. 2008; 7, 235-243.
71. Montiglio P, Garant D, Thomas D, Réale D. Individual variation in temporal activity patterns in open-field tests. Anim. Behav. 2010; 80, 905-912.
72. Kalueff AV, Keisala T, Minasyan A, Kuuslahti M, Tuohimaa P. Temporal stability of novelty exploration in mice exposed to different open field tests. Behav. Proc. 2006; 72, 104-112.
73. Hofer MA, Shair HN, Brunelli SA. Ultrasonic vocalizations in rat and mouse pups. Curr. Protoc. Neurosci. 2002; Chapter 8; Unit 8.14.
74. Vanderschuren LJMJ, Achterberg EJM, Trezza V. The neurobiology of social play and its rewarding value in rats. Neurosci. Biobehav. Rev. 2016; 70, 86-105.
75. Martin P, Kraemer HC. Individual differences in behavior and their statistical consequences. Anim. Behav. 1987; 35 (5), 1366-1375.
76. Bate S, Clark R. Experimental Design. In: The Design and Statistical analysis of Animal Experiments, Cambridge UK: Cambridge University Press, pp. 30-121.
77. Van der Staay FJ. Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. Brain Res. Rev. 2006; 52, 131-159.
78. Rodgers RJ, Cao BJ, Dalvi A, Holmes A. Animal models of anxiety: an ethological perspective. Braz. J. Med. Biol. Res. 1997; 30, 289-304.
79. Fuchs E, Flügge G. Experimental animal models for the simulation of depression and anxiety. Dialogues Clin. Neurosci. 2006; 8 (3), 323-333.
80. Bushby EV, Friel M, Goold C, Gray H, Smith L, Collins LM. Factors influencing individual variation in farm animal cognition and how to account for these statistically. Front. Vet. Sci. 2018; 5, 193.

81. Cleasby IR, Nakagawa S, Schielzeth H. Quantifying the predictability of behavior: statistical approaches for the study of between-individual variation and the within-individual variance. Methods Ecol. Evol. 2015; 6, 27-37.
82. Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, et al. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 2018; 6, e4794.
83. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens HH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 2009; 24 (3), 127-135.
84. Van Driel KS, Talling JC. Familiarity increases consistency in animal tests. Behav. Brain Res. 2005; 159, 243-245.
85. Gouveia K, Hurst JL. Optimizing reliability of mouse performance in behavioral testing: the major role of non-aversive handling. Sci. Rep. 2017; 7, 44999.
86. Lewejohann L, Reinhard C, Schrewe A, Brandewiede J, Haemisch A, Görtz N, et al. Environmental bias? Effects of housing conditions, laboratory environment and experimenter on behavioral tests. Genes Brain Behav. 2006; 5, 64-72.
87. Bohlen M, Hayes ER, Bohlen B, Bailoo BD, Crabbe JC, Wahlsten D. Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. Behav. Brain Res. 2014; 217, 46-54.
88. Spruijt BM, de Visser L. Advanced behavioral screening: automated home cage ethology. Drug Discov. Today, 2006; 3, 231-237.
89. Kaufman AB, Rosenthal R. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Anim. Behav. 2009; 78 (6), 1478-1491.
90. Salllinen J, Link RE, Haapalinna A, Viitamaa T, Kulatunga M, Sjöholm B, et al. Genetic alteration of alpha 2C-adrenoceptor expression in mice: influence on locomotor, hypothermic, and neurochemical effects of dexmedetomidine, a subtype-nonselective alpha 2-adrenoceptor agonist. Mol. Pharmacol. 1997; 51, 36-46.
91. Votava M, Hess L, Slíva J, Kršiak M, Agová V. Dexmedetomidine selectively suppresses dominant behavior in aggressive and sociable mice. Eur. J. Pharmacol. 2005; 523, 79-85.
92. Fairbanks CA, Kitto KF, Nguyen HO, Stone LS, Wilcox GL. Clonidine and dexmedetomidine produce antinociceptive synergy in mouse spinal cord. Anesthesiology 2009; 110, 648-647.
93. Wilkinson M, Manchester EL. Strain differences in brain alpha2 and beta-adrenergic receptor binding in dystrophic mice. Brain Res. Bull. 1983; 11, 743-745.

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

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

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

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

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

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

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

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

(I) Dimension		**Estimate ± SEM**	$t_{(df)}$	***P***	**Cohens *d* [95% CI]**
Avoidance					
Trial 1 vs 5					
	A	-0.726 ± 0.96	$-7.593_{(708)}$	**< 0.0001**	-1.015 [-1.283, -0.747]
	B	0.985 ± 0.96	$10.288_{(708)}$	**< 0.0001**	1.377 [1.105, 1.650]
Exploration					
Trial 1 vs 5					
	A	0.085 ± 0.05	$1.583_{(708)}$	0.1138	0.319 [-0.077, 0.716]
	B	-0.953 ± 0.06	$-15.274_{(708)}$	**< 0.0001**	-3.588 [-4.085, -3.090]
Locomotion (rank transformed)					
Trial 1 vs 5					
	A	98.06 ± 23.1	$4.250_{(708)}$	**< 0.0001**	0.498 [0.267, 0.730]
	B	-252.68 ± 32.6	$-7.740_{(708)}$	**< 0.0001**	-1.285 [-1.618, -0.952]
(II) Dimension		**Estimate ± SEM**	$t_{(df)}$	***P***	**Cohen's *d* [95% CI]**
Avoidance					
A vs B	Trial 1	-0.828 ± 0.111	$-7.430_{(177)}$	**< 0.0001**	-1.158 [-1.488, -0.827]
	Trial 2	-0.578 ± 0.111	$-5.186_{(177)}$	**< 0.0001**	-0.808 [-1.127, -0.489]
	Trial 3	0.173 ± 0.111	$1.554_{(177)}$	0.1220	0.242 [-0.066, 0.551]
	Trial 4	0.634 ± 0.111	$5.691_{(177)}$	**< 0.0001**	0.887 [0.565, 1.208]
	Trial 5	0.883 ± 0.111	$7.926_{(177)}$	**< 0.0001**	1.235 [0.901, 1.569]
Exploration					
A vs B	Trial 1	0.190 ± 0.057	$3.318_{(177)}$	**0.0011****	0.715 [0.283, 1.147]
	Trial 2	0.129 ± 0.078	$1.658_{(177)}$	0.0992	0.485 [-0.095, 1.066]
	Trial 3	-0.240 ± 0.083	$-2.900_{(177)}$	**0.0042****	-0.903 [-1.525, -0.281]
	Trial 4	-0.633 ± 0.095	$-6.658_{(177)}$	**< 0.0001**	-2.384 [-3.134, -1.635]
	Trial 5	-0.848 ± 0.100	$-8.452_{(177)}$	**< 0.0001**	-3.192 [-4.009, -2.375]
Locomotion (rank transformed)					
A vs B	Trial 1	43.7 ± 46.0	$0.949_{(177)}$	0.3438	0.222 [-0.240, 0.685]
	Trial 2	-152.1 ± 35.7	$-4.260_{(177)}$	**< 0.0001**	-0.774 [-1.140, -0.406]
	Trial 3	-227.4 ± 33.6	$-6.774_{(177)}$	**< 0.0001**	-1.157 [-1.510, -0.798]
	Trial 4	-272.1 ± 31.8	$-8.551_{(177)}$	**< 0.0001**	-1.384 [-1.730, -1.033]
	Trial 5	-307.1 ± 31.2	$-9.851_{(177)}$	**< 0.0001**	-1.562 [-1.910, -1.208]

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

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

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

(a) Balanced and unbalanced pool combined: post hoc comparisons					
Dimension		**Estimate ± SEM**	**z**	**P**	**Cohens d [95% CI]**
Avoidance					
Strain effect	129S2 vs C	-0.192 ± 0.205	-0.940	0.3471	-0.239 [-0.738, 0.260]
	129S2 vs B6N	-0.465 ± 0.213	-2.184	0.0290	-0.577 [-1.104, -0.050]
	C vs B6N	-0.273 ± 0.204	-1.335	0.1819	-0.338 [-0.838, 0.162]
Exploration					
Strain effect	129S2 vs C	-0.220 ± 0.140	-1.571	0.1161	-0.399 [-0.901, 0.103]
	129S2 vs B6N	-0.730 ± 0.145	-5.024	**< 0.0001**	-1.327 [-1.891, -0.763]
	C vs B6N	-0.511 ± 0.139	-3.663	**0.0002**	-0.928 [-1.499, -0.408]
Locomotion (rank transformed)					
Strain effect	129S2 vs C	-15.8 ± 5.28	-2.994	**0.0028**	-0.760 [-1.270, -0.246]
	129S2 vs B6N	-41.6 ± 5.50	-7.563	**< 0.0001**	-2.000 [-2.620, -1.381]
	C vs B6N	-25.8 ± 5.27	-4.886	**< 0.0001**	-1.240 [-1.780, -0.700]

(b) Balanced pool only: post hoc comparisons					
Dimension		**Estimate ± SEM**	**z**	**P**	**Cohens d [95% CI]**
Avoidance					
Strain effect	129S2 vs C	-0.235 ± 0.316	-0.744	0.4568	-0.317 [-1.155, 0.521]
	129S2 vs B6N	-0.377 ± 0.333	-1.130	0.2587	-0.507 [-1.396, 0.382]
	C vs B6N	-0.141 ± 0.264	-0.535	0.5927	-0.190 [-0.889, 0.509]
Exploration					
Strain effect	129S2 vs C	-0.352 ± 0.225	-1.563	0.1181	-0.655 [-1.520, 0.185]
	129S2 vs B6N	-0.909 ± 0.237	-3.828	**0.0001**	-1.719 [-2.690, -0.743]
	C vs B6N	-0.557 ± 0.188	-2.961	**0.0031**	-1.053 [-1.800, -0.310]
Locomotion (rank transformed)					
Strain effect	129S2 vs C	-9.86 ± 4.51	-2.183	0.0290	-0.93 [-1.79, -0.06]
	129S2 vs B6N	-23.83 ± 4.76	-5.006	**< 0.0001**	-2.25 [-3.29, -1.209]
	C vs B6N	-13.98 ± 3.77	-3.702	**0.0002**	-1.32 [-2.09, -0.550]

(c) Unbalanced pool only: post hoc comparisons					
Dimension		**Estimate ± SEM**	**z**	**P**	**Cohens d [95% CI]**
Avoidance					
Strain effect					
	129S2 vs C	-0.088 ± 0.331	-0.267	0.7893	-0.096 [-0.805, 0.612]
	129S2 vs B6N	-0.495 ± 0.324	-1.528	0.1266	-0.540 [-1.246, 0.165]
	C vs B6N	-0.407 ± 0.333	-1.223	0.2215	-0.444 [-1.163, 0.276]
Exploration					
Strain x Treatment interaction					
Treatment (T)	129S2 vs C	-0.012 ± 0.219	0.056	0.9556	-0.028 [-0.961, 1.018]
	129S2 vs B6N	-0.153 ± 0.217	-0.708	0.4793	-0.354 [-1.338, 0.630]
	C vs B6N	-0.166 ± 0.220	-0.753	0.4514	-0.382 [-1.381, 0.617]
Control (C)	129S2 vs C	-0.273 ± 0.219	-1.245	0.2133	-0.630 [-1.634, 0.374]
	129S2 vs B6N	-1.097 ± 0.217	-5.062	**< 0.0001**	-2.531 [-3.691, -1.371]
	C vs B6N	-0.824 ± 0.219	-3.757	**0.0002**	-1.901 [-2.997, -0.805]
C versus T	129S2	-0.031 ± 0.217	-0.144	0.8852	-0.072 [-1.052, 0.908]
	C	0.254 ± 0.217	1.171	0.2415	0.586 [-0.405, 1.576]
	B6N	0.913 ± 0.217	4.207	**< 0.0001**	2.105 [0.997, 3.213]
Locomotion (rank transformed)					
Strain effect	129S2 vs C	-6.36 ± 3.85	-1.653	0.0983	-0.597 [-1.32, 0.126]
	129S2 vs B6N	-18.26 ± 3.77	-4.845	**< 0.0001**	-1.713 [-2.52, -0.903]
	C vs B6N	-11.89 ± 3.87	-3.076	**0.0021**	-1.116 [-1.88, -0.354]

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

			Design: balanced			Design: not balanced		
Dimension	**Experimenter**	**Strain/ treatment**	**mean**	**ci_lower**	**ci_upper**	**mean**	**ci_lower**	**ci_upper**
Avoidance behavior	Exp I	129S2 -saline	0.023019	-1.61607	1.662108	0.200246	-1.38651	1.787001
		129S2-dex	0.129693	-1.43332	1.692703	0.016335	-1.94767	1.980338
		C-saline	0.101698	-1.47017	1.673563	-0.03734	-1.26342	1.188748
		C-dex	0.441862	-0.82784	1.711569	0.193076	-1.60137	1.987524
		B6N-saline	0.429354	-0.13032	0.989028	0.09654	-0.59096	0.784039
		B6N-dex	0.625128	-0.28483	1.535088	0.571633	-0.73361	1.876876
	Exp II	129S2 -saline	-0.69476	-1.12455	-0.26498	-0.37338	-1.97587	1.229113
		129S2-dex	-0.73457	-1.73569	0.266541	-0.59157	-2.28593	1.102792
		C-saline	-0.55749	-1.06447	-0.05052	-0.5685	-1.24379	0.10679
		C-dex	0.316467	-1.34841	1.981347	-0.08268	-2.19269	2.027338
		B6N-saline	-0.27818	-0.78015	0.223796	-0.24184	-0.77131	0.287622
		B6N-dex	0.197786	-1.37734	1.772915	0.817472	0.487122	1.147822
Exploration	Exp I	129S2 -saline	-0.39387	-0.87417	0.086425	-0.4229	-0.89838	0.052583
		129S2-dex	-0.50552	-0.85873	-0.1523	-0.43842	-0.92738	0.050539
		C-saline	-0.25477	-0.88672	0.377188	-0.222	-0.76146	0.317457
		C-dex	-0.33452	-0.79876	0.129717	-0.32847	-0.7302	0.07327
		B6N-saline	0.187644	-0.33855	0.713839	0.767557	-0.60851	2.143629
		B6N-dex	0.242245	-0.85694	1.341427	-0.23107	-0.678	0.21587
	Exp II	129S2 -saline	-0.16467	-0.31148	-0.01785	-0.13366	-0.59896	0.33163
		129S2-dex	-0.24404	-0.48126	-0.00681	-0.05553	-0.21117	0.100119
		C-saline	0.352888	-0.74396	1.449732	0.287712	-0.26056	0.835983
		C-dex	-0.1803	-1.0436	0.682995	-0.12176	-1.29527	1.051738
		B6N-saline	0.8984	-0.08546	1.882258	0.870524	0.261631	1.479417
		B6N-dex	0.396512	-1.26627	2.05929	0.028014	-0.48706	0.543084
Locomotion	Exp I	129S2 -saline	-0.30246	-0.84295	0.238036	-0.03936	-0.55794	0.479215
		129S2-dex	-1.14415	-3.43442	1.146118	-1.3907	-3.63961	0.858205
		C-saline	-0.14697	-1.39312	1.099179	0.038583	-0.8603	0.937462
		C-dex	-0.21732	-2.33342	1.898771	0.025043	-0.884	0.934084
		B6N-saline	0.851491	0.214075	1.488907	0.57257	0.083174	1.061966
		B6N-dex	0.473804	-0.11642	1.064024	-0.39373	-2.17827	1.390809
	Exp II	129S2 -saline	-0.35352	-1.28484	0.57779	-0.11184	-0.58074	0.357052
		129S2-dex	-0.25024	-0.53849	0.038012	-0.06912	-0.55721	0.418971
		C-saline	-0.01579	-0.37472	0.343136	0.194522	-0.21018	0.599226
		C-dex	0.012291	-0.74927	0.773856	-0.34784	-0.83812	0.142431
		B6N-saline	0.875258	0.393329	1.357187	1.045723	-0.19156	2.283011
		B6N-dex	0.217619	-0.45245	0.887684	0.476164	0.132941	0.819387

Box II

Summary of results of the behavioral and physiological data from Phase I of Chapter 4.

M.H. van der Goot

Summary of results of between strain differences of the data collected for Phase I of Chapter 4. Differences analyzed for avoidance behavior, risk assessment, arousal, exploration, locomotion and plasma corticosterone (pCORT) levels. Strains compared: BALB/cAnNCrl (C, *n* = 59), C57BL/6NCrl (B6N, *n* = 60) and 129S2/SvPasCrl (129S2, *n* = 60). Circulating corticosterone levels (pCORT) were assessed at three time points for each individual. The first sample was collected one week prior to the behavioral test (7 days ± 1). The second sample was collected directly after behavioral testing, approximately 35 minutes after the first mHB trial. Finally, the third sample was taken a week after behavioral testing (7 days ± 1). Within each mouse, all samples were collected on approximately the same the time of day to avoid fluctuation of pCORT due to circadian rhythm (Spencer and Deak, 2017).

Between strain differences in behavior analyzed with linear mixed models (LMMs) using a 3 (strain) x 2 (experimenter) x 5 (trial) mixed factorial design. Strain, experimenter and trial were included as fixed factors, including all interactions. Individual mouse (ID), slope (trial nested in ID), batch and test order were included as random effects. Between strain differences in pCORT levels were analyzed with a generalized least squares model (GLS using a 3 (strain) x 3 (sampling time) mixed factorial design. Strain, sampling time and were included as fixed factors, while day of test was included as fixed covariate. The variable pCORT was logarithmically transformed to achieve normality of the residuals.

I. *Post hoc* comparisons trial 1 vs 5, per experimenter (Dunn-Šidák adjusted α = 0.016921): **(a)** Experimenter I: decrease in C, increase in 129S2, stable in B6N (C, *very large* effect size, d = 1.702, 95% CI [1.159, 2.244], $t_{(692)}$ = 6.248, P < 0.0001; 129S2, *large* effect size, d = -1.025, 95% CI [-1.693, -0.356], $t_{(692)}$ = -3.020, P – 0.0026; B6N, *small* effect size, d = 0.153, 95% CI [-0.232, 0.539], $t_{(692)}$ = 0.781, P = 0.4351). **(b)** Experimenter II: decrease in C, increase in 129S2, decrease in B6N (C, *very large* effect size, d = 1.775, 95% CI [1.241, 2.309], $t_{(692)}$ = 6.628, P < 0.0001; 129S2, *very large* effect size, d = -3.116, 95% CI [-3.802, -2.430], $t_{(692)}$ = -9.182, P < 0.0001; B6N, *medium* effect size, d = 0.572, 95% CI [0.186, 0.958], $t_{(692)}$ = 2.195, P = 0.0037).

II. *Post hoc* comparisons between strain on each trial, per experimenter (Dunn-Šidák adjusted α = 0.016921 for comparisons on trials 1 and 5; adjusted α = 0.025321 for comparisons on trials 2, 3 and 4): **(a)** C *vs* 129S2: Experimenter I: Trial 1, C > 129S2; Trial 4 and 5, C < 129S2 (Trial 1, *medium* effect size,

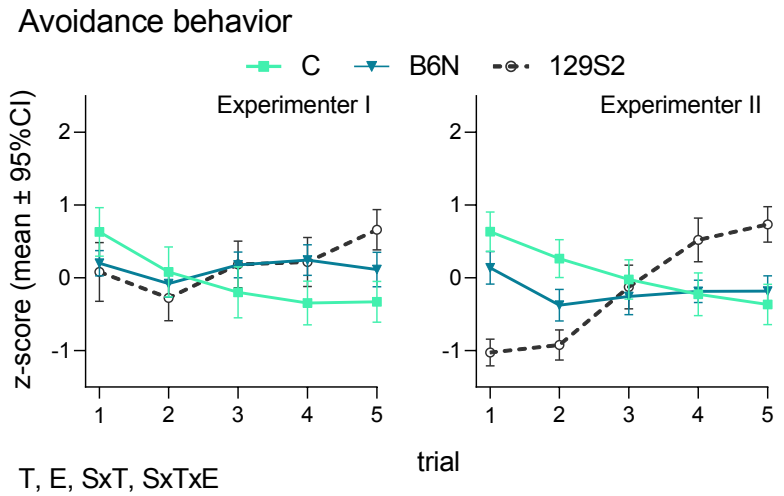


Figure 1. Avoidance behavior trajectories of C, B6N and 129S2, as scored by each experimenter. Behavior expressed as integrated behavioral z-scores. Results are presented as means with 95% CI. Effects were significant in a LMMs at P < 0.05. S: significant main effect of strain; T: significant main effect of trial; SxT: significant interaction between strain and trial; SxTxE: significant three-way interaction between strain, trial and experimenter. **Results LMM** Avoidance behavior: trial effect: $F_{(4,692)}$ = 18.96, P < 0.0001; experimenter effect: $F_{(1, 173)}$ = 5.56, P = 0.0194; interaction strain x trial: $F_{(8, 692)}$ = 23.04, P < 0.0001; interaction strain x trial x experimenter: $F_{(8, 692)}$ = 4.96, P < 0.0001).

d = -0.972, 95% CI [-1.690, -0.254], $t_{(173)}$ = -2.701, P = 0.0004; Trial 4, *medium* effect size, d = 0.997, 95% CI [0.276, 1.719], $t_{(173)}$ = 2.758, P = 0.0064; Trial 5, *very large* effect size, d = 1.754, 95% CI [1.006, 2.502], $t_{(173)}$ = 4.781, P < 0.0001). Experimenter II: Trial 1 and 2, C > 129S2; Trial 4 and 5, C < 129S2 (Trial 1, *very large* effect size, d = -2.940, 95% CI [-3.711, -2.169], $t_{(173)}$ = -8.228, P < 0.0001; Trial 2, *very large* effect size, d = -2.103, 95% CI [-2.840, -1.367], $t_{(173)}$ = -5.915, P < 0.0001; Trial 4, *large* effect size, d = 1.325, 95% CI [0.603, 2.048], $t_{(173)}$ = 3.691, P = 0.0003; Trial 5, *very large* effect size, d = 1.950, 95% CI [1.202, 2.699], $t_{(173)}$ = 5.353, P < 0.0001). **(b)** C *vs* B6N: Experimenter I: Trial 1, C > B6N; Trial 4 and 5, C < B6N (Trial 1, *medium* effect size, d = 0.759, 95% CI [0.161, 1.357], $t_{(173)}$ = 2.527, P = 0.0124; Trial 4, *large* effect size, d = -1.042, 95% CI [-1.650, 0.435], $t_{(173)}$ = -3.446, P = 0.0007; Trial 5, *medium* effect size, d = -0.789, 95% CI [-1.405, -0.174], $t_{(173)}$ = -2.556, P = 0.0114). Experimenter II: Trial 1 and 2, C > B6N (Trial 1, *medium* effect size, d = 0.879, 95% CI [0.285, 1.474], $t_{(173)}$ = 2.959, P = 0.0035; Trial 2, *large* effect size, d = 1.133, 95% CI

[0.538, 1.728], $t_{(173)}$ = 3.837, P = 0.0002). **(c)** 129S2 vs B6N: Experimenter I: Trial 5, 129S2 > B6N (Trial 5, *medium* effect size, d = 0.965, 95% CI [0.286, 1.644], $t_{(173)}$ = 2.837, P = 0.0051). Experimenter II: Trial 1 and 2, 129S2 < B6N; Trial 4 and 5, 129S2 > B6N (Trial 1, *very large* effect size, d = -2.061, 95% CI [-2.753, -1.369], $t_{(173)}$ = -6.194, P < 0.0001; Trial 2, *medium* effect size, d = -0.971, 95% CI [-1.632, -0.310], $t_{(173)}$ = -2.934, P = 0.0038; Trial 4, *large* effect size, d = 1.255, 95% CI [0.581, 1.929], $t_{(173)}$ = 3.750, P = 0.0002; Trial 5, *very large* effect size, d = 1.627, 95% CI [0.934, 2.321], $t_{(173)}$ = 4.783, P < 0.0001).

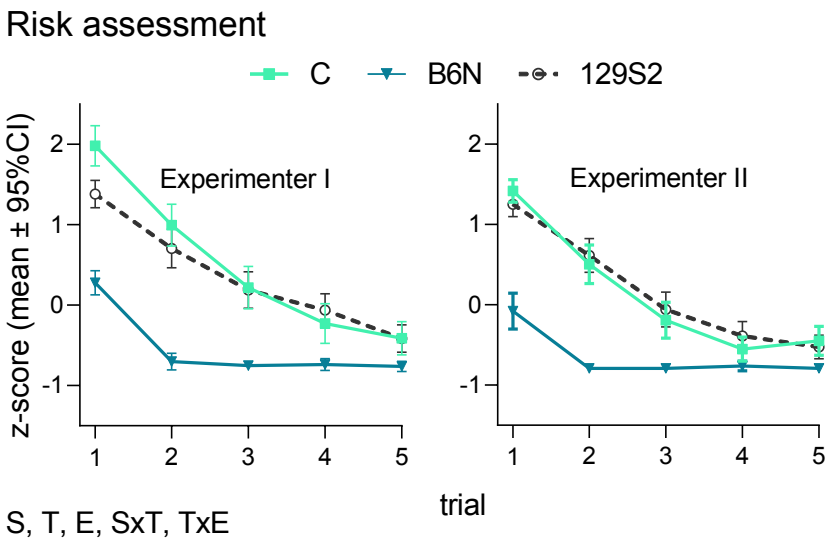


Figure 2. Risk assessment trajectories of C, B6N and 129S2, as scored by each experimenter. Behavior expressed as integrated behavioral z-scores. Results are presented as means with 95% CI. Effects were significant in a LMM at P < 0.05. S: significant main effect of strain; T: significant main effect of trial; SxT: significant interaction between strain and trial; E: significant main effect of experimenter; TxE: significant interaction between trial and experimenter. **Results LMM** Risk assessment: strain effect: $F_{(2,173)}$ = 261.66, P < 0.0001; trial effect: $F_{(4, 692)}$ = 310.42, P < 0.0001; interaction strain x trial: $F_{(8, 692)}$ = 31.09, P < 0.0001; experimenter effect: $F_{(1,173)}$ = 20.97, P < 0.0001; interaction trial x experimenter: $F_{(4,173)}$ = 3.99, P = 0.0033).

I. *Post hoc* comparisons trial 1 vs 5, across experimenters (Dunn-Šidák adjusted α = 0.016921): Decrease in C, 129S2 and B6N: (C, *very large* effect size, d = 3.619, 95% CI [3.210, 4.028], $t_{(692)}$ = 21.020, P < 0.0001; B6N, *large* effect size, d = 1.490, 95% CI [1.305, 1.676], $t_{(692)}$ = 12.465, P < 0.0001; 129S2, *very large* effect size, d = 3.038, 95% CI [2.701, 3.376], $t_{(692)}$ = 23.585, P < 0.0001).

II. *Post hoc* comparisons between strain on each trial, across experimenters (Dunn-Šidák adjusted α = 0.016921 for comparisons on trials 1 and 5; adjusted α = 0.025321 for comparisons on trials 2, 3 and 4): **(a)** C *vs* 129S2: Trial 1, C > 129S2 (Trial 1, *medium* effect size, d = -0.651, 95% CI [-0.996, -0.305], $t_{(173)}$ = -3.794, P = 0.0002). **(b)** C *vs* B6N: Trials 1-5, C > B6N (Trial 1, *very large* effect size, d = 2.712, 95% CI [2.308, 3.125], $t_{(173)}$ = 18.571, P < 0.0001; Trial 2, *very large* effect size, d = 2.543, 95% CI [2.147, 2.938], $t_{(173)}$ = 17.383, P < 0.0001; Trial 3, *large* effect size, d = 1.335, 95% CI [1.013, 1.657], $t_{(173)}$ = 9.126, P < 0.0001; Trial 4, *medium* effect size, d = 0.614, 95% CI [0.318, 0.910], $t_{(173)}$ = 4.195, P < 0.0001; Trial 5, *medium* effect size, d = 0.588, 95% CI [0.292, 0.883], $t_{(173)}$ = 4.018, P < 0.0001). **(c)** 129S2 *vs* B6N: Trials 1-5, 129S2 > B6N (Trial 1, *very large* effect size, d = 2.066, 95% CI [1.735, 2.396], $t_{(173)}$ = 16.489, P < 0.0001; Trial 2, *very large* effect size, d = 2.390, 95% CI [2.036, 2.745], $t_{(173)}$ = 19.080, P < 0.0001; Trial 3, *large* effect size, d = 1.421, 95% CI [1.131, 1.710], $t_{(173)}$ = 11.340, P < 0.0001; Trial 4, *medium* effect size, d = 0.892, 95% CI [0.626, 1.157], $t_{(173)}$ = 7.118, P < 0.0001; Trial 5, *medium* effect size, d = 0.517, 95% CI [0.264, 0.771], $t_{(173)}$ = 4.130, P < 0.0001).

III. *Post hoc* comparisons between experimenters on each trial, across strains (Dunn-Šidák adjusted α = 0.025321 for comparisons on trials 1 and 5; α = 0.05 for comparisons on trials 2, 3 and 4): Trial 1-4, Exp I > Exp II (Trial 1, *medium* effect size, d = 0.593, 95% CI [0.351, 0.835], $t_{(173)}$ = 4.877, P < 0.0001; Trial 2, *small* effect size, d = 0.379, 95% CI [0.138, 0.619], $t_{(173)}$ = 3.113, P = 0.0022; Trial 3, *small* effect size, d = 0.393, 95% CI [0.152, 0.634], $t_{(173)}$ = 3.232, P = 0.0015; Trial 4, *small* effect size, d = 0.381, 95% CI [0.140, 0.621], $t_{(173)}$ = 3.130, P = 0.0021).

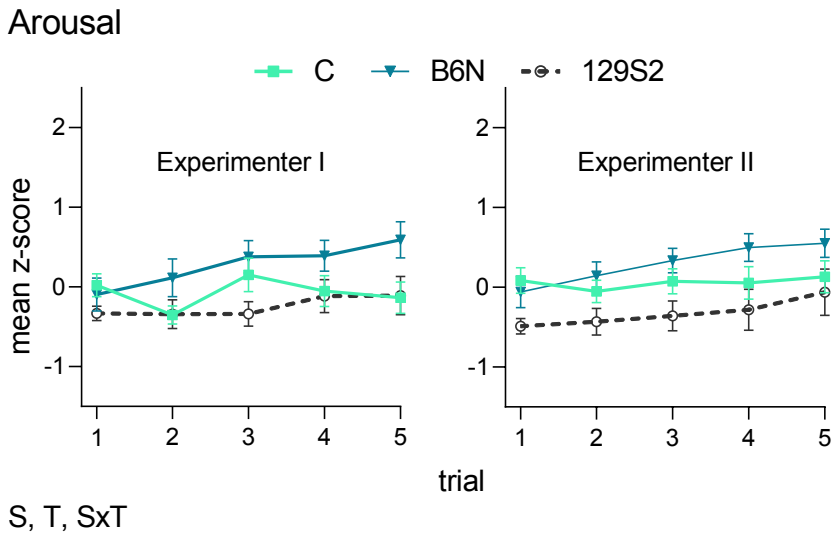


Figure 3. Arousal trajectories of strains C, B6N and 129S2, as scored by each experimenter. Behavior expressed as integrated behavioral z-scores. Results are presented as means with 95% CI. Effects were significant in a LMM at $P < 0.05$. S: significant main effect of strain; T: significant main effect of trial; SxT: significant interaction between strain and trial. **Results LMM** Arousal: strain effect: $F_{(2,173)} = 40.14$, $P < 0.0001$; trial effect: $F_{(4, 692)} = 12.09$, $P < 0.0001$; interaction strain x trial: $F_{(8, 692)} = 5.28$, $P < 0.0001$.

I) *Post hoc* comparisons trial 1 vs 5, across experimenters (Dunn-Šidák adjusted α = 0.016921): Stable in C, increase in 129S2 and B6N (C, *small* effect size, $d = 0.125$, 95% CI [-0.307, 0.559], $t_{(652)} = 0.570$, $P = 0.0001$; 129S2, *medium* effect size, $d = -0.766$, 95% CI [-1.201, -0.323], $t_{(652)} = -3.481$, $P = 0.0005$; B6N, *very large* effect size, $d = -1.595$, 95% CI [-1.922, -1.037], $t_{(652)} = -6.680$, $P < 0.0001$).

II. *Post hoc* comparisons between strain on each trial, across experimenters (Dunn-Šidák adjusted α = 0.016921 for comparisons on trials 1 and 5; adjusted α = 0.025321 for comparisons on trials 2, 3 and 4): **(a)** C *vs* 129S2: Trials 1-3, C > 129S2 (Trial 1, *large* effect size, $d = -1.083$, 95% CI [-1.469, -0.697], $t_{(173)} = -5.796$, $P < 0.0001$; Trial 2, *medium* effect size, $d = -0.428$, 95% CI [-0.821, -0.036], $t_{(173)} = -2.167$, $P = 0.0316$; Trial 3, *large* effect size, $d = -1.016$, 95% CI [-1.452, -0.580], $t_{(173)} = -4.747$, $P < 0.0001$). **(b)** C *vs* B6N: Trials 2-5, C < B6N (Trial 2, *medium* effect size, $d = -0.777$, 95% CI [-1.173, -0.381], $t_{(173)} = -3.963$, $P = 0.0001$; Trial 3, *medium* effect size, $d = -0.579$, 95% CI [-0.999, -0.158],

$t_{(173)} = -2.745$, $P = 0.0067$; Trial 4, *large* effect size, $d = -1.009$, 95% CI [-1.490, -0.528], $t_{(173)} = -4.246$, $P < 0.0001$; Trial 5, *large* effect size, $d = -1.310$, 95% CI [-1.849, -0.771], $t_{(173)} = -4.967$, $P < 0.0001$). **(c)** 129S2 *vs* B6N: Trials 1-5, 129S2 < B6N (Trial 1, *medium* effect size, $d = -0.788$, 95% CI [-1.167, -0.408], $t_{(173)} = -4.201$, $P < 0.0001$; Trial 2, *large* effect size, $d = -1.205$, 95% CI [-1.616, -0.795], $t_{(173)} = -6.615$, $P < 0.0001$; Trial 3, *very large* effect size, $d = -1.595$, 95% CI [-2.046, -1.143], $t_{(173)} = -7.526$, $P < 0.0001$; Trial 4, *large* effect size, $d = -1.452$, 95% CI [-1.938, -0.967], $t_{(173)} = -6.227$, $P < 0.0001$; Trial 5, *very large* effect size, $d = -1.501$, 95% CI [-2.045, -0.956], $t_{(173)} = -5.686$, $P < 0.0001$).

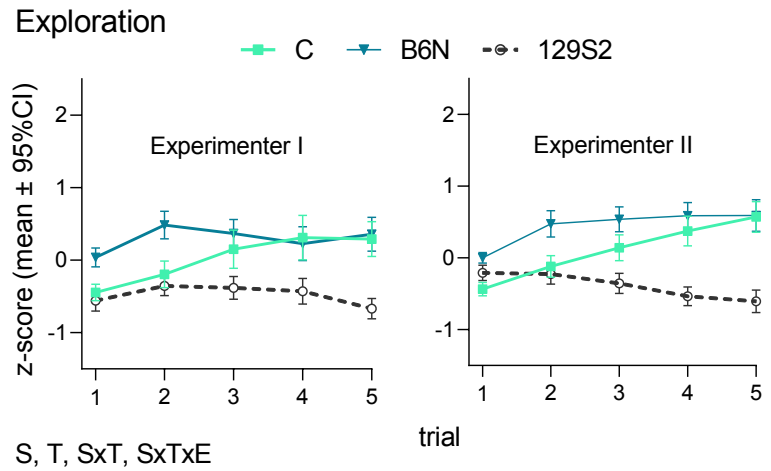


Figure 4. Exploration trajectories of strains C, B6N and 129S2, as scored by each experimenter. Behavior expressed as integrated behavioral z-scores. Results are presented as means with 95% CI. Effects were significant in a LMMs at $P < 0.05$. S: significant main effect of strain; T: significant main effect of trial; SxT: significant interaction between strain and trial; SxTxE: significant three-way interaction between strain, trial and experimenter. **Results LMM** Exploration: strain effect: $F_{(2,173)} = 53.54$, $P < 0.0001$; trial effect: $F_{(4,692)} = 41.14$, $P < 0.0001$; interaction strain x trial: $F_{(8, 692)} = 19.54$, $P < 0.0001$; interaction strain x trial x experimenter: $F_{(8, 692)} = 2.96$, $P = 0.0029$.

I. *Post hoc* comparisons trial 1 vs 5, per experimenter (Dunn-Šidák adjusted α = 0.016921): **(a)** Experimenter I: Increase in C and B6N, stable in 129S2 (C, *very large* effect size, $d = -3.474$, 95% CI [-4.404, 2.544], $t_{(692)} = -7.494$, $P < 0.0001$; 129S2, *medium* effect size, $d = 0.518$, 95% CI [-0.379, 1.415], $t_{(692)} = 1.134$, $P = 0.2572$; B6N, *very large* effect size, $d = -1.515$, 95% CI [-2.584, -0.446], $t_{(692)} = -2.793$, $P = 0.0054$). **(b)** Experimenter II: Increase in C and B6N, decrease

in 129S2 (C, *very large* effect size, d = -4.763 95% CI [-5.694, -3.831], $t_{(692)}$ = -10.449, P < 0.0001; 129S2, *very large* effect size, d = 1.867, 95% CI [0.964, 2.770], $t_{(692)}$ = 4.088, P < 0.0001; B6N, *very large* effect size, d = -2.781, 95% CI [-3.857, -1.704], $t_{(692)}$ = -5.125, P < 0.0001).

II. *Post hoc* comparisons between strain on each trial, per experimenter (Dunn-Šidák adjusted α = 0.016921 for comparisons on trials 1 and 5; adjusted α = 0.025321 for comparisons on trials 2, 3 and 4): **(a)** C *vs* 129S2: Experimenter I: Trials 3-5, C > 129S2 (Trial 3, *very large* effect size, d = -2.525, 95% CI [-3.711, -1.341], $t_{(173)}$ = -4.318, P < 0.0001; Trial 4, *very large* effect size, d = -3.510, 95% CI [-4.878, -2.142], $t_{(173)}$ = -5.262, P < 0.0001; Trial 5, *very large* effect size, d = -4.521, 95% [-5.858, -3.185], $t_{(173)}$ = -7.155, P < 0.0001).

Experimenter II: Trial 1, C < 129S2, Trials 3-5, C > 129S2 (Trial 1, *large* effect size, d = 1.071, 95% CI [0.291, 1.852], $t_{(173)}$ = 2.740, P = 0.0068; Trial 3, *very large* effect size, d = -2.353, 95% CI [-3.524, -1.184], $t_{(173)}$ = -4.064, P = 0.0001; Trial 4, *very large* effect size, d = -4.299, 95% CI [-5.680, -2.918], $t_{(173)}$ = -6.509, P < 0.0001; Trial 5, *very large* effect size, d = -5.559, 95% CI [-6.928, -4.190], $t_{(173)}$ = -8.880, P < 0.0001). **(b)** C *vs* B6N: Experimenter I: Trial 1 and 2, C < B6N (Trial1, *very large* effect size, d = -2.277, 95% CI [-2.995, -1.560], $t_{(692)}$ = -6.653, P < 0.0001; Trial 2, *very large* effect size, d = -3.210, 95% CI [-4.383, -2.039], $t_{(692)}$ = -5.653, P < 0.0001). Experimenter II: Trial 1-3, C < B6N (Trial 1, *very large* effect size, d = -2.066, 95% CI [-2.771, -1.361], $t_{(692)}$ = -6.084, P < 0.0001; Trial 2, *very large* effect size, d = -2.800, 95% CI [-3.953, -1.648], $t_{(692)}$ = -4.963, P < 0.0001; Trial 3, *very large* effect size, d = -1.874, 95% CI [-3.083, -0.667], $t_{(692)}$ = -3.105, P = 0.0022). **(c)** 129S2 vs B6N:

Experimenter I: Trials 1-5, 129S2 < B6N (Trial 1, *very large* effect size, d = -2.806, 95% CI [-3.650, -1.964], $t_{(173)}$ = -7.023, P < 0.0001; Trial 2, *very large* effect size, d = -3.974, 95% CI [-5.178, -2.770], $t_{(173)}$ = -6.955, P < 0.0001; Trial 3, *very large* effect size, d = -3.552, 95% CI [-4.712, -2.393], $t_{(173)}$ = -6.395, P < 0.0001; Trial 4, *very large* effect size, d = -3.108, 95% CI [-4.419, -1.797], $t_{(173)}$ = -4.834, P < 0.0001; Trial 5, *very large* effect size, d = -4.840, 95% CI [-6.293, -3.388], $t_{(173)}$ = -7.033, P < 0.0001). Experimenter II: Trials 2-5, 129S2 < B6N (Trial 2, *very large* effect size, d = -3.312, 95% CI [-4.494, -2.131], $t_{(173)}$ = -5.797, P < 0.0001; Trial 3, *very large* effect size, d = -4.228, 95% CI [-5.413, -3.044], $t_{(173)}$ = -7.613, P < 0.0001; Trial 4, *very large* effect size, d = -5.314, 95% CI [-6.703, -3.926], $t_{(173)}$ = -8.266, P < 0.0001; Trial 5, *very large* effect size, d = -5.643, 95% [-7.128, -4.158], $t_{(173)}$ = -8.199, P < 0.0001).

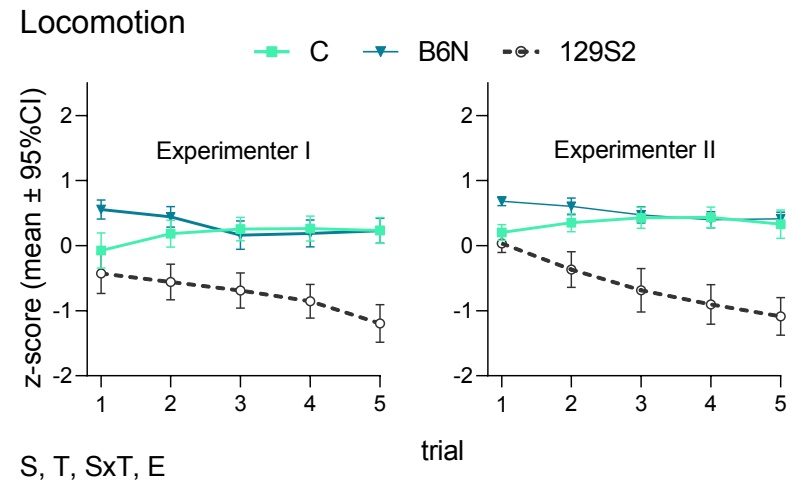


Figure 5. Locomotion trajectories of strains C, B6N and 129S2, as scored by each experimenter. Behavior expressed as integrated behavioral z-scores. Results are presented as means with 95% CI. Effects were significant in a LMM at P < 0.05. S: significant main effect of strain; T: significant main effect of trial; SxT: significant interaction between strain and trial; E: significant main effect of experimenter; **Results LMM** Locomotion: strain effect: $F_{(2,173)}$ = 111.12, P < 0.0001; trial effect: $F_{(4,692)}$ = 15.51, P < 0.0001; interaction strain x trial: $F_{(8, 692)}$ = 16.85, P < 0.0001; experimenter effect: $F_{(1,173)}$ = 11.00, P = 0.0011.

I. *Post hoc* comparisons trial 1 vs 5, across experimenters (Dunn-Šidák adjusted α = 0.016921): Increase in C, decrease in 129S2 and B6N (C, *medium* effect size, d = -0.561, 95% CI [-0.933, -0.190], $t_{(692)}$ = -2.977, P = 0.0030; 129S2, *very large* effect size, d = 2.428, 95% CI [1.854, 3.003], $t_{(692)}$ = 8.520, P < 0.0001; B6N, *large* effect size, d = 1.479, 95% CI [0.448, 1.078], $t_{(692)}$ = 4.801, P < 0.0001).

II. *Post hoc* comparisons between strain on each trial, across experimenters (Dunn-Šidák adjusted α = 0.016921 for comparisons on trials 1 and 5; adjusted α = 0.025321 for comparisons on trials 2, 3 and 4): **(a)** C *vs* 129S2: Trials 2-5, C > 129S2 (Trial 2, *very large* effect size, d = -1.875, 95% CI [-2.449, -1.302], $t_{(173)}$ = -6.885, P < 0.0001; Trial 3, *very large* effect size, d = -2.643, 95% CI [-3.265, -2.022], $t_{(173)}$ = -9.409, P < 0.0001; Trial 4, *very large* effect size, d = -3.152, 95% CI [-3.816, -2.488], $t_{(173)}$ = -10.848, P < 0.0001; Trial 5, *very large* effect size, d = -3.658, 95% CI [-4.368, -2.948], $t_{(173)}$ = -12.149, P < 0.0001). **(b)** C *vs* B6N: Trial 1 and 2, C < B6N (Trial 1, *large* effect size, d = -1.426, 95% CI [-1.854,

-0.998], $t_{(173)}$ = -7.033, P < 0.0001; Trial 2, *medium* effect size, d = -0.653, 95% CI [-1.078, -0.228], $t_{(173)}$ = -3.075, P = 0.0025). **(c)** 129S2 *vs* B6N: Trials 1-5, 129S2 < B6N (Trial 1, *very large* effect size, d = -2.093, 95% CI [-2.643, -1.544], $t_{(173)}$ = -8.228, P < 0.0001; Trial 2, *very large* effect size, d = -2.528, 95% CI [-3.112, -1.946], $t_{(173)}$ = -9.644, P < 0.0001; Trial 3, *very large* effect size, d = -2.582, 95% CI [-3.184, -1.982], $t_{(173)}$ = -9.528, P < 0.0001; Trial 4, *very large* effect size, d = -3.009, 95% CI [-3.649, -2.370], $t_{(173)}$ = -10.713, P < 0.0001; Trial 5, *very large* effect size, d = -3.758, 95% CI [-4.459, -3.058], $t_{(173)}$ = -12.887, P < 0.0001).

III. *Post hoc* comparisons between experimenters, across strain and trial (α = 0.05): Exp I < Exp II (*small* effect size, d = -0.446, 95% CI [-0.772, -0.121], $t_{(173)}$ = -2.738, P = 0.0068).

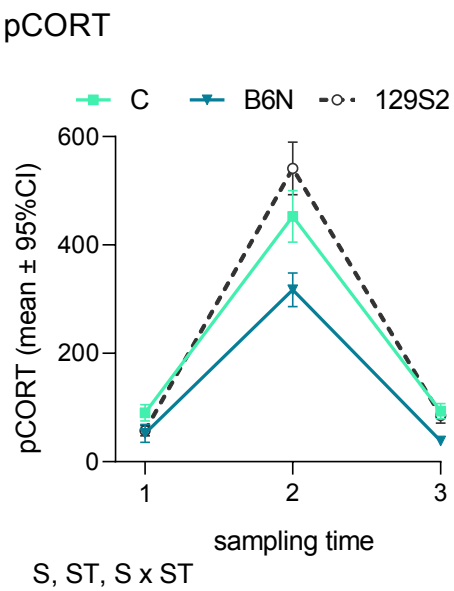


Figure 6. Blood plasma corticosterone (pCORT) levels in strains C, B6N and 129S2 one week prior to behavioral test (sampling time 1), directly after mHB test (sampling time 2) and one week after behavioral test (sampling time 3). Results expressed as nmol/L and presented as means with 95% CI. Effects were significant in a LMM at P < 0.05. S: significant main effect of strain; ST: significant main effect of sampling time; S x ST: significant interaction between strain and sampling time. **Results GLS** pCORT: strain effect: $F_{(2,504)}$ = 62.34, P < 0.0001; sampling time effect: $F_{(2,504)}$ = 870.02, P < 0.0001; interaction strain x sampling time: $F_{(4, 504)}$ = 7.22, P < 0.0001.

I) *Post hoc* comparisons between sampling times for each strain separately (Dunn-Šidák adjusted α = 0.012741): **(a)** Sampling time 1 *vs* 2: Increase in C, 129S2 and B6N (C, *very large* effect size, d = -2.271, 95% CI [-2.603, -1.940], $t_{(240)}$ = -15.621, P < 0.0001; B6N, *very large* effect size, d = -2.670, 95% CI [-3.015, -2.325], $t_{(244)}$ = -18.584, P < 0.0001; 129S2, *very large* effect size, d = -3.056, 95% CI [-3.424, -2.688], $t_{(238)}$ = -20.723, P < 0.0001). **(b)** Sampling time 2 *vs* 3: Decrease in C, 129S2 and B6N (C, *very large* effect size, d = 2.028, 95% CI [1.728, 2.328], $t_{(248)}$ = 15.372, P < 0.0001; B6N, *very large* effect size, d = 2.782, 95% CI [2.448, 3.115], $t_{(250)}$ = 20.821, P < 0.0001; 129S2, *very large* effect size, d = 2.422, 95% CI [2.109, 2.735], $t_{(253)}$ = 18.554, P < 0.0001). **(c)** Sampling time 1 *vs* 3: Increase in 129S2, stable in C and B6N (C, *small* effect size, d = -0.243, 95% CI [-0.598, 0.112], $t_{(354)}$ = -1.348, P = 0.3695; B6N, *small* effect size, d = 0.112, 95% CI [-0.242, 0.465], $t_{(356)}$ = 0.662, P = 0.8084; 129S2, *medium* effect size, d = -0.634, 95% CI [-0.992, -0.276], $t_{(350)}$ = -3.513, P = 0.0015).

II. *Post hoc* comparisons between strain on each sampling time (Dunn-Šidák adjusted α = 0.012741): **(a)** C *vs* 129S2: sampling time 2, C < 129S2 (*small* effect size, d = 0.244, 95% CI [0.081, 0.407], $t_{(170)}$ = 2.995, P = 0.0088). **(b)** C *vs* B6N: sampling time 1, 2 and 3, C > B6N (sampling time 1, *medium* effect size, d = 0.839, 95% CI [0.460, 1.216], $t_{(173)}$ = 4.509, P < 0.0001; sampling time 2, *small* effect size, d = 0.440, 95% CI [0.271, 0.608], $t_{(170)}$ = 5.381, P < 0.0001; sampling time 3, *large* effect size, d = 1.193, 95% CI [0.838, 1.548], $t_{(168)}$ = 7.126, P < 0.0001). **(c)** 129S2 *vs* B6N: sampling time 2 and 3, 129S2 > B6N (sampling time 2, *medium* effect size, d = 0.684, 95% CI [0.506, 0.862], $t_{(170)}$ = 8.326, P < 0.0001; sampling time 3, *large* effect size, d = 1.043, 95% CI [0.697, 1.390], $t_{(168)}$ = 6.285, P < 0.0001).

Reference

Spencer, R.L., Deak, T., 2017. A users guide to HPA axis research. Physiol. Behav. 178, 43-65, https://doi.org/10.1016/j.physbeh.2016.11.014.

Chapter 5

Chromosomal assignment of quantitative trait loci influencing the change of anxiety-related modified Hole Board behavior in male laboratory mice using B6.A-consomics

In preparation

Marloes H. van der Goot[1,2,] Marijke C. Laarakker[1], Maaike Labots[1], Saskia S. Arndt[1], Hein A. van Lith[1,2]

[1] Section Animals in Science and Society, Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands

[2] Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands

Abstract

Background: Anxiety-related behavior in laboratory mice is a complex behavioral construct that is expressed by both anxiety-related and activity behavior. Chromosome substitution strains (CSSs, also called consomic strains or lines) have proven a powerful tool in the search for quantitative trait loci (QTLs) that modulate the expression of this behavior. Chromosomes 10, 15 and 19 have come forward as prominent regions modulating anxiety-related behavior expressed in the modified Hole Board (mHB). In these studies, the expression of anxiety is often measured as the response to a stressor. We now conducted a consomic strain survey on the temporal aspect of this behavior: the change of anxiety-related behavior over time.

Method: Previously collected data of a male B6.A CSS panel – including the A/J (A) donor strain and C57BL/6J (B6) host strain – was re-examined by transforming behavioral variables measured in a single mHB trial to trajectories of five 60-second epochs. Behavioral responses of CSSs were characterized using the multivariate approach of integrated behavioral z-scoring. Selection of CSSs was based on a combination of effect size measures and statistical significance.

Results: In line with our previous analyses on behavioral response in this mHB trial, mouse chromosomes 10 and 19 came forward as regions that modulated the expression of anxiety-related behavior across time in this panel. This effect remained when controlling for locomotor activity in the consomic line with chromosome 19 from A, but not in the CSS with chromosome 10 from A.

Conclusion: Mouse chromosome 19 not only plays a role in the expression of anxiety-related behavior in the mHB, but also modulates the temporal component of this behavior.

Keywords

Anxiety, habituation, chromosome substitution strain, A/J, C57BL/6J, behavioral trajectory, area under the curve.

1. Introduction

Anxiety-related behavior in laboratory mice is a complex behavioral construct that involves the display of both anxiety-related and activity behavior (Belzung and Griebel, 2001; Ohl, 2003). From a genetic perspective, research focusing on the genetic underpinnings of this construct shows that anxiety-related behavior in mice has a complex inheritance, in which multiple genes interact with each other as well as with epigenetic and environmental factors (Clément et al., 2002).

In this field, the identification of candidate genes of anxiety-related behavior starts with the mapping of QTLs (Baud and Flint, 2017): the most likely region(s) of a chromosome that is/are associated with genetic variation for a particular trait (Lander and Botstein, 1989; Labots et al., 2016). One of the available specialized sources that helps to detect QTLs are CSSs (Nadeau et al., 2012). CSSs are created by transferring a single chromosome from one inbred strain – the donor strain – onto the genetic background of another inbred strain – the host strain – by repeated backcrossing (Singer et al., 2004; Nadeau et al., 2012). The identification of chromosomes that contain at least one QTL for a particular phenotype is then carried out through comparison of that phenotype of each consomic line with the host strain. In this procedure, the complex genome is partitioned in a defined and reproducible manner. This makes CSSs highly suitable for complex trait analysis (Nadeau et al., 2012) and a powerful starting tool for detecting meaningful QTLs (de Mooij-van Malsen et al., 2009). The first complete mouse CSSs panel ever produced was created from B6 and A strains (Nadeau et al., 2000). These two inbred strains are frequently used in anxiety research because of their contrasting anxiety phenotypes (e.g. Trullas and Skolnick, 1993; van Gaalen and Steckler, 2000; Bouwknecht and Paylor, 2002; see Laarakker et al., 2011 for an overview of inbred mouse strain surveys ranking A and B6 on anxiety-related behavior). B6 mice are known for their low anxious, highly active behavioral profile (Rogers et al., 1999; Bolivar et al., 2000; Garner, 2005; Tam and Cheung, 2020). A mice are typically characterized by high levels of anxiety-related behavior and low locomotor activity (Trullas and Skolnick, 1993; Bolivar et al., 2000; Bothe et al., 2005; Moy et al., 2007; Kas et al., 2008; Laarakker et al., 2011; Tam and Cheung, 2020).

Unconditoned anxiety-related behavior is often assessed in behavioral paradigms such as the open field test (OF), the light-dark box (LD), the elevated plus maze (EPM) and the mHB (Ohl, 2005). When applying these paradigms

while using B6 and A as parental strains in different mapping populations, QTL analyses revealed multiple chromosomal regions associated with anxiety-related behavior (Mathis et al., 1995; Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Singer et al., 2004; Gill and Boyle, 2005; Singer et al., 2005; Zhang et al., 2005; de Mooij-van Malsen et al., 2009). Interestingly, a number of these studies also associated QTLs with the temporal component of anxiety-related behavior (i.e. the change of anxiety-related behavior across time: Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Zhang et al., 2005; de Mooij-van Malsen et al., 2009). This phenomenon is of special interest in preclinical anxiety research that studies the adaptive quality of anxiety-related responses in mice.

In rodents, exposure to a novel environment or stimulus induces a biologically adaptive anxiety response that enables individuals to respond to potential threat (Ohl, 2003). A decrease of this response during repeated or prolonged exposure is then considered an adaptive emotional response that allows individuals to adapt to environmental challenges (Ohl et al., 2008). The adaptive quality of anxiety responses is thus assessed by means of a decrease in anxiety-related behavior, or lack thereof (Boleij et al., 2012; Salomons et al., 2010a; Salomons et al. 2010b; Salomons et al. 2010c; Salomons et al., 2013). In an adaptive phenotype, such decrease is typically combined with an increase in activity (van der Goot et al., 2020). In a non-adaptive phenotype, initial levels of anxiety fail to decrease (i.e. increase or remain unchanged) while activity levels decrease or remain stable (van der Goot et al., 2020).

Adaptive capacities have indeed also been shown to differ between B6 and A. In B6, initial low levels of avoidance behavior are typically combined with high initial locomotor and exploratory activity. Both anxiety-related and activity behavior decrease during repeated or prolonged exposure to novelty (Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Bolivar et al., 2000; Bothe et al., 2005). The temporal profile of A mice is more ambiguous, with reports of both an increase in avoidance behavior (Bothe et al., 2005) or high levels of avoidance behavior that remained stable over time (Gershenfeld and Paul, 1997; Bolivar et al., 2000) during exposure to an OF. Similarly, activity behavior has been described as remaining stable (Logue et al., 1997; Bothe et al., 2005), increasing (Bolivar et al., 2000) or decreasing after repeated exposure to the OF (Gershenfeld et al., 1997).

The few QTL analyses incorporating the temporal component of anxiety-related behavior suggest that adaptive capacities of anxiety responses may also have a heritable component, with a prominent role for mouse chromosomes 1, 10 and 19 (Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Zhang et al., 2005; de Mooij-van Malsen et al., 2009; de Mooij-van Malsen et al., 2013). These studies however were either based on intercrosses of B6 and A strains that were assessed in the OF test (Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Zhang et al., 2005), or assessed in a panel of female B6.A CSSs and a subsequent intercross of B6 and a consomic line in the home cage, rather than in response to novelty (de Mooij-van Malsen et al., 2009; de Mooij-van Malsen et al., 2013). To our knowledge no chromosomal QTL assignment regarding the change of anxiety-related behavior over time during a five minute trial has been reported in a panel of male B6.A CSSs in response to a novel stimulus in an unconditioned test of anxiety. In the present study we therefore conducted a consomic strain survey on the temporal component of anxiety-related behavior, using previously collected and partly published (Laarakker et al., 2008; Labots et al., 2016) mHB data of the host, donor and consomic strains (see Table 1).

The mHB combines the characteristics of an OF, a hole board and a LD (Ohl et al., 2001; Labots et al., 2015) and as such allows for the expression of a range of anxiety-related and activity behaviors. It is therefore especially suitable for phenotyping of complex behavioral constructs, such as anxiety-related behavior (Ohl et al., 2001). In the original dataset, the host (B6), donor (A) and the panel of CSSs were subjected to a single five minute mHB trial. The results of this single trial data were published in Laarakker et al., (2008) and Labots et al., (2016) and showed that anxiety-related behavior was higher in A than B6, while activity behavior was lower in A. This CSS survey revealed a prominent role for mouse chromosomes 10, 15 and 19 in the display of anxiety-related behavior, and to a lesser extent chromosomes 1, 6, X and Y (Laarakker et al., 2008; Labots et al., 2016).

To assess the temporal component of anxiety-related behavior in these data, we transformed the raw data to trajectories of five 60-second epochs. The consomic strain survey subsequently compared the trajectories and the area under the curve (AUC) between the host and donor strain, and between the host strain and the panel of consomic lines.

In order to do this, we first summarized the separate behavioral variables. It has been shown that multivariate analysis increases the power to detect meaningful QTLs, in comparison to univariate analysis (Turri et al., 2004). Previous mHB studies showed that anxiety-related, but also activity variables observed in the mHB are related to each other (Laarakker et al., 2008; Labots et al., 2016). Labots et al., (2016; 2018) demonstrated that these behaviors can be summarized to five behavioral dimensions (avoidance behavior, risk assessment, arousal, exploration and locomotion) through the approach of integrated behavioral z-scoring. Furthermore, Labots et al., (2016) recommended the use of effect size measures combined with statistical significance to improve the selection method of consomic mouse strains, as statistical significance (represented by *P* values) in itself does not necessarily predicate practical importance (Lovell, 2013; Festing, 2014). Our present survey thus compared (integrated) behavioral z-scores of these five behavioral dimensions between host and donor or consomic lines. In addition, meaningful differences between host and donor, and between host and consomic line were based on a combination of effect sizes and statistical significance. Given the literature reporting a role of chromosomes 10 and 19, and the results of the single trial data, we hypothesized a role for QTL on chromosomes 10 and 19.

2. Materials and Methods

2.1. General

The present study is reported according to the revised ARRIVE guidelines (ARRIVE 2.0; https://www.nc3rs.org.uk/revision-arrive-guidelines (Percie du Sert et al., 2020a; 2020b).

2.2. Animals and housing

Our consomic strain survey was performed on previously collected data, for the most part of which was previously published (Laarakker et al., 2008; Labots et al., 2016). Table 1 presents an overview of which data was published and which not. The original experiment was performed on naive male mice of the following inbred strains: A (the donor strain, n = 30), B6 (the host strain, n = 27) and a set of CSSs between these parental strain (n = 6 per consomic line). The official nomenclature of the consomic lines is C57BL/6J-Chr #$^{A/J}$/NaJ, which is simplified to CSS-# (# = mouse chromosome number/letter) in the rest of this report. Unfortunately, digital raw behavioral mHB data for CSS-4 was not available due to technical reasons. Thus, in total twenty consomic lines were used. We tested more host strain animals compared to consomic mice to improve power to detect a chromosome that contains a QTL. According to Belknap (2003) a ratio close to 4.5:1 is the most efficient for selecting CSSs that contain a QTL. Animals were purchased from the Jackson Laboratory (Bar Harbor, ME, USA). Charles River Netherlands (Maastricht, the Netherlands) coordinated the shipping from the Jackson Laboratory to the testing facility in Utrecht.

All mice were 4-6 weeks old upon arrival. Animals were housed at the Central Laboratory Animal Research Facility of Utrecht University (location 'Paviljoen'). Testing took place in the same room as where the animals were housed, test equipment had been installed in the room prior to arrival of the animals. Mice were housed individually in Macrolon® Type II-L cages (size: 365 x 207 x 140 mm, floor area 530 cm^2, Techniplast, Milan, Italy) with standard bedding material (autoclaved Aspen Chips, Abedd-Dominik Mayr KEG, Köflach, Austria) and a tissue (KLEENEX® Facial Tissue, Kimberley-Clark Professional BV, Ede, the Netherlands), a cardboard shelter (Technilab-BMI BV, Someren, the Netherlands) and handful of paper shreds (EnviroDri, Technilab-BMI BV) as enrichment. Food (CRM, Expanded, Special Diets Services Essex, UK) and tap water were available *ad libitum*. All mice were kept in a sound attenuated laboratory animal housing room for a habituation period of at least two weeks under a reversed light-dark schedule (white light: 7PM-7AM, maximal 150 lux; red light: 7AM-7PM, maximal

5 lux). Relative humidity was kept constant at approximately 50 ± 5%. Average room temperature was maintained at 21 ± 2 ℃, with a ventilation rate of 15-20 air changes per hour. A radio played constantly as background noise. During the habituation period animals were handled at least four times per week by the person performing the behavioral tests (MCL). Handling included picking up the mice at the tail base and placing it on the hand or arm of the experimenter, or by restraining it for a few seconds at random times of the day.

2.3. Behavioral testing

Animals were tested at the age of 6-10 weeks. Behavioral assessment occurred in the mHB. This apparatus has been described extensively elsewhere (Ohl et al., 2001; Labots et al., 2015) and will only briefly described here. The mHB consists of a grey PVC opaque box (100 x 50 x 50 cm, length x width x height) with a board made of the same material (60 x 20 x 0.5 cm, length x width x height) functioning as an unprotected area, as it is positioned in the center of box. The board stacks 23 cylinders (3 x 3 cm, diameter x height) in three lines. The area around the board is divided into 10 rectangles (20 x 15 cm) and 2 squares (20 x 20 cm). The periphery was illuminated with red light (1-5 lux) and functioned as the protected area. The central board was illuminated with an additional red light lamp (80 W, 35 lux on the board) to increase the aversive nature of the central (unprotected) area in comparison to the box.

The experimental protocol has previously been described in detail (Laarakker et al., 2008). At the start of the behavioral test, mice were placed in the mHB and freely explored the set-up for 5 minutes, while their behavior was scored by a trained observer (MCL). Testing occurred during the active phase of the animals, between 10AM and 2PM, and all behavioral tests were videotaped for raw data storage. The behavioral variables were scored live using the computer software Observer 4.1 (Noldus, Wageningen, the Netherlands). Between behavioral tests, feces and urine were removed from the set-up and the experimental compartment was cleaned with a damp tissue.

2.4. Behavioral variables

As described in the Introduction, we used two approaches to analyze the temporal component of anxiety-related behavior: *i)* analysis of the change of behavior over a five-minute epoch (trajectories), and *ii)* analysis of the AUCs combined with latencies, which is often used as a summary of the magnitude of a behavioral response. For both approaches, the raw Observer-output of each behavioral variable (except for latencies), for each individual mouse, was first transposed to trajectories of five 60-second epochs.

Table 1. Characteristics of male mice included in the present consomic strain survey.

Mouse strain name	Abbreviated Name	JAX stock number	*n*	Published[a]
C57BL/6J	B6	000664	6	Yes
C57BL/6J	B6	000664	21	No
A/J	A	000646	30	Yes
C57BL/6J-Chr 1$^{A/J}$/NaJ	CSS-1	004379	6	Yes
C57BL/6J-Chr 2$^{A/J}$/NaJ	CSS-2	004380	6	Yes
C57BL/6J-Chr 3$^{A/J}$/NaJ	CSS-3	004381	6	Yes
C57BL/6J-Chr 5$^{A/J}$/NaJ	CSS-5	004383	6	Yes
C57BL/6J-Chr 6$^{A/J}$/NaJ	CSS-6	004384	6	Yes
C57BL/6J-Chr 7$^{A/J}$/NaJ	CSS-7	004385	6	Yes
C57BL/6J-Chr 8$^{A/J}$/NaJ	CSS-8	004386	6	Yes
C57BL/6J-Chr 9$^{A/J}$/NaJ	CSS-9	004387	6	Yes
C57BL/6J-Chr 10$^{A/J}$/NaJ	CSS-10	004388	6	Yes
C57BL/6J-Chr 11$^{A/J}$/NaJ	CSS-11	004389	6	Yes
C57BL/6J-Chr 12$^{A/J}$/NaJ	CSS-12	004390	6	Yes
C57BL/6J-Chr 13$^{A/J}$/NaJ	CSS-13	004391	6	Yes
C57BL/6J-Chr 14$^{A/J}$/NaJ	CSS-14	004392	6	Yes
C57BL/6J-Chr 15$^{A/J}$/NaJ	CSS-15	004393	6	Yes
C57BL/6J-Chr 16$^{A/J}$/NaJ	CSS-16	004394	6	Yes
C57BL/6J-Chr 17$^{A/J}$/NaJ	CSS-17	004395	6	Yes
C57BL/6J-Chr 18$^{A/J}$/NaJ	CSS-18	004396	6	Yes
C57BL/6J-Chr 19$^{A/J}$/NaJ	CSS-19	004397	6	Yes
C57BL/6J-Chr X$^{A/J}$/NaJ	CSS-X	004398	6	Yes
C57BL/6J-Chr Y$^{A/J}$/NaJ	CSS-Y	004399	6	Yes

[a]Data have been published in Laarakker et al., (2008) and Labots et al., (2016).

Labots et al., (2018) demonstrated that the behavioral variables observed in the mHB can be summarized to integrated behavioral z-scores for five behavioral dimensions: avoidance behavior, risk assessment, arousal, exploration and locomotion. In addition, the behavioral dimensions avoidance behavior, risk assessment and arousal may be summarized to the integrated z-score for the motivational system 'anxiety'. The method of integrated behavioral z-scoring was first proposed by Guilloux et al., (2011) and extended and improved by Labots et al., (2018) as a multidimensional approach for behavioral phenotyping in mice.

Labots et al., (2016) showed that the integrated behavioral z-scores of these five dimensions and the motivational system 'anxiety' correlated significantly with the orthogonal factors that resulted from a factor analysis that was performed on the same variables by Laarakker et al., (2008). Furthermore, the range of heritability (h^2) of these integrated z-scores was deemed acceptable for behavioral phenotypes in mice, and high enough to select consomic strains (Labots et al., 2016). In the present survey thus, we summarized the separate mHB variables to integrated behavioral z-scores for further analysis. The exact procedure is described extensively in Labots et al., (2018). In short, it entails that behavioral variables that measure different aspects of the same behavioral dimension are normalized and combined to a single score representing that particular behavioral dimension or motivational system. First, normalization of each variable is done by z-score transformation, which measures the amount of standard deviations each observation is above or below the mean of a reference group (Labots et al., 2018). In the present study we used the pooled data of all included strains and epochs (except for latencies) as the reference population + reference condition, as suggested by Labots et al., (2018).

Second, the transformed separate variables are averaged within each behavioral dimension, resulting in an integrated behavioral z-score. The integrated behavioral z-score for the anxiety motivational system is then subsequently obtained by averaging the integrated z-scores for avoidance behavior, risk assessment and arousal. Table 2 presents an overview of the separate variables observed in the mHB, and which variables were attributed to which behavioral dimension in this study.

The exact composition of integrated z-scores differed between the trajectory and the AUC analyses combined with latencies. Table 2 lists the original behavioral variables scored by Laarakker et al., (2008), and how these were

summarized to integrated z-scores for the trajectories (3rd column), and the AUC variables combined with latencies (4th column). The integrated behavioral z-scores for the trajectory analyses included the frequencies (*number of times* a behavior was performed per 60 second epoch) and duration (*percentage of time* per 60-second epoch a behavior was performed). It is common to include the *latencies* of each behavior (the first time a behavior is displayed) in formation of these integrated z-scores (Labots et al., 2016; Labots et al., 2018). For the trajectory data these values however could not be included in composition of the integrated z-scores because latency variables could not be translated to five 60-second epochs. This posed a problem for the composition of the integrated z-scores for risk assessment and locomotion, as these dimensions are composed of two variable types: a frequency variable and a latency variable (see Table 2). The trajectory of risk assessment therefore was solely based on the z-score of 'total number of stretched attends', and the trajectory of locomotion was solely based on the z-score of 'total number of line crossings' (Table 2). In the remainder of this paper, the trajectories of these z-transformed variables are referred to as the trajectories for risk assessment and locomotion.

For the AUC analyses, the areas under the curve were computed for the frequency and duration for each of the five epochs of each behavioral variable (0-60 s, 61-120 s, etc.), using the trapezoid method. For each frequency or duration variable, the sum of these five AUC values (i.e. one triangle and four trapezoids) formed the total AUC for that variable. These single-variable AUC scores were supplemented with the latency variables to form integrated behavioral z-scores for each behavioral dimension (Table 2).

2.5. Statistical analysis

All analyses were conducted with R version 4.0.0 in R-Studio (R Core Team, 2020). All Figures were created with GraphPad Prism (GraphPad Prism version 7.04 for Windows, Graphpad Software, La Jolla, California USA, www.graphpad.com).

Table 2. List of behavioral variables measured in the mHB and used in this study.

Motivational system/ Behavioral dimension	Behavioral variable scored in the mHB	Trajectories[1]	AUCs combined with latencies[1]
Anxiety-related behavior			
Avoidance behavior	Total number of board entries	-z	-z
	Latency until first board entry		z
	Percentage of time spent on the board	-z	-z
Risk assessment	Total number of stretched attends	z	z
	Latency until first stretched attend		-z
Arousal	Total number of self-groomings	z	z
	Latency until the first self-grooming		-z
	Percentage of time spent on self-grooming	z	z
	Total number of boli	z	z
	Latency until first boli is produced		-z
Exploration	Total number of rearings in the box	z	z
	Latency until first rearing in the box		-z
	Total number of rearings on the board	z	z
	Latency until first rearing on the board		-z
	Total number of hole explorations	z	z
	Latency until first hole exploration		-z
	Total number of hole visits	z	z
	Latency until first hole visit		-z
Locomotion	Total number of line crossings	z	z
	Latency until first line crossing		-z

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

2.5.1. Trajectory analysis

Trajectories of the (integrated) z-scores for anxiety, avoidance behavior, risk assessment, arousal, exploration and locomotion were analyzed with generalized linear mixed models (GLMM) using the package 'glmmTMB' (Brooks et al., 2017). For each motivational system/behavioral dimension, GLMMs compared the behavioral response over time between B6 and the donor strain (A) or consomic line, resulting in 21 comparisons per system/dimension.

Strain differences were analyzed using a strain (2) x epoch (5) mixed factorial design, with repeated measures on the second factor. Strain and epoch were included as fixed factors including their interaction, while time of day and season at test were included as fixed covariates. Individual mouse (ID) was included as random effect. Additional analyses were conducted for the motivational system 'anxiety'. In these models the z-score for locomotion or exploration was included as a covariate to control for a potentially confounding effect of locomotion or exploration on anxiety. All models were run with a heterogeneous Toeplitz covariance structure. Model assumptions were assessed using the package 'DHARMa', which uses a simulation based approach for residual diagnostics of multi-level/mixed regression models (Hartig, 2020). Avoidance behavior was rank transformed in the model comparing B6 (host) and A (donor strain) to improve the residual distribution. The integrated z-score for exploration was log-transformed ($y = \text{Log}[x + 1]$) in all comparisons to meet the assumptions, with the exception of the model comparing exploration in B6 and A, in which exploration was rank transformed. The models for the (integrated) z-scores of risk assessment and arousal included a correction for zero-inflation on the fixed factors by means of a hurdle model because the residuals were zero-inflated.

Main and interaction effects from all GLMMs were derived using Wald Chi Square tests with corresponding *P* value. Due to the large number of analyses per motivational system/behavioral dimension ($n = 21$), we employed a significance threshold of $P < 0.004$ and a suggestive threshold of $0.004 \leq P < 0.05$. The threshold for significance was suggested by Belknap (2003) as an acceptable cutoff when comparing donor and consomic lines with a host strain. As described in the introduction, our objective was to base the detection of meaningful differences between host and donor, and host and consomic line, on a combination of *P* values and effect sizes. Unfortunately, commonly applied effect sizes as (partial) eta squared cannot be computed for individual model terms due to the way the variance is partitioned in GLMMs (Rights and Sterba, 2019). The coefficient of determination (R^2) may be used as a summary statistic that describes the part of the variance that is explained by the model (Nakagawa and Schielzeth, 2013), but this statistic does not apply to individual model terms.

We therefore followed up on main and interaction effects by *post hoc* tests, and used the Cohen's *d* effect size that followed from these tests to classify differences between host and donor/host and consomic line as suggestive/significant/no evidence for a meaningful effect (see below).

Post hoc pairwise comparisons were conducted using the package 'emmeans' (Lenth, 2020). To reduce the probability of a Type I error due to multiple comparisons, the α was adjusted using a Dunn-Šidák correction in *post hoc* tests, using the formula $\alpha = 1 - [1 - 0.004]^{1/\lambda}$ where λ is the number of times a data set is used in multiple comparisons. A significant/suggestive main effect of epoch was followed up by *post hoc* contrasts between the first and the last epoch (to assess the change of behavior over time). A significant/suggestive main effect of strain (without any interaction) was followed up by *post hoc* contrasts of between strain differences on each epoch (Table 3). A significant/suggestive interaction between strain and epoch was followed up by *post hoc* comparison of the first and the last epoch within each strain (i.e. donor or consomic), and comparison between strains (i.e host *vs.* donor or consomic) on each separate epoch. Table 3 presents an overview of the adjusted α-value for each comparison and contrast. All *post hoc* tests were summarized as beta-estimates and their corresponding standard error, *t* statistic and *P* values.

Effect sizes for *post hoc* tests were reported as Cohen's *d* with 95% CI, and obtained via the package 'emmeans'. The guidelines provided by Wahlsten (2011) were used to interpret the absolute values of Cohen's *d* (|*d*|). This review of various phenotypes proposed the following interpretation of effects for neurobehavioral mouse studies: *small* effect, $|d| \leq 0.5$; *medium* effect, $0.5 < |d| < 1.0$; *large* effect, $1.0 \leq |d| < 1.5$; *very large* effect, $|d| \geq 1.5$. We used a combination of *P* values (GLMMs) and Cohen's *d* effect sizes to determine evidence for a meaningful difference between host and donor strain, and between host strain and the consomic lines. *Large/very large* effect sizes of which the 95% CI contained the value zero were not included for evaluation as this amounts to a non-significant ($P \geq 0.05$) result (Nakagawa and Cuthill, 2007). A very large chromosomal effect (Cohen's $d \geq 1.5$) combined with $P < 0.004$ for strain and/or an interaction between strain and epoch, was considered indicative of significant evidence for that chromosome harboring a QTL. A Cohen's $d < 1.0$ and $P > 0.05$ were considered as no evidence for a chromosome harboring a QTL. All other threshold combinations were regarded as indicative of suggestive evidence (see Table 4). In *post hoc* tests between/across the five epochs, the largest Cohen's *d* effect size of all epoch-related comparisons/contrasts was used to determine the evidence for a meaningful effect.

Table 3. *Post hoc*-comparisons trajectory analysis: adjusted significance tresholds.[a]

Results section	GLMM effect		γ	Suggestive threshold	Significant threshold
trajectories	Strain (S) /Strain x Epoch (S x E)	A/Cons-E1 *vs.* B6-E1	2	$0.002002 \leq P < 0.025321$	$P < 0.002002$
		A-Cons-E2 *vs.* B6-E2	1	$0.004 \leq P < 0.05$	$P < 0.004$
		A-Cons-E3 *vs.* B6-E3	1	$0.004 \leq P < 0.05$	$P < 0.004$
		A-Cons-E4 *vs.* B6-E4	1	$0.004 \leq P < 0.05$	$P < 0.004$
		A-Cons-E5 *vs.* B6-E5	2	$0.002002 \leq P < 0.025321$	$P < 0.002002$
	Strain x Epoch (S x E)	A/Cons-E1 *vs.* A/Cons-E5	2	$0.002002 \leq P < 0.025321$	$P < 0.002002$
		B6-E1 *vs.* B6-E5	2	$0.002002 \leq P < 0.025321$	$P < 0.002002$
	Epoch (E)	E1 *vs.* E5	1	$0.004 \leq P < 0.05$	$P < 0.004$

[a]Suggestive/significant thresholds corrected for multiple comparisons using a Dunn-Šidák correction, in the case of a suggestive/significant main effect of strain (S), epoch (E) or interaction between strain and epoch (S x E). γ = the number of times a data set is used in a *post hoc* comparison. A = donor strain, Cons = consomic line, B6 = host strain, E = epoch.

Table 4. Type of evidence for the GLMMs, based on Cohen's *d* and *P*-values.

Omnibus test		***Post hoc* comparison**	**Evidence for a meaningful QTL/ Meaningful difference between host and donor strain**
Strain effect (S)	Interaction Strain x Epoch	Epoch 1, 2, 3, 4, 5[a]	
$P > 0.05$	$P > 0.05$	$\|d\| < 1.0$	No
$P > 0.05$	$P > 0.05$	$1.0 \leq \|d\| < 1.5$ (#)[b]	No
$P > 0.05$	$P > 0.05$	$\|d\| \geq 1.5$ (##)	No
$P > 0.05$	$0.004 > P \leq 0.05$ (*)	$\|d\| < 1.0$	No
$P > 0.05$	$0.004 > P \leq 0.05$ (*)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P > 0.05$	$0.004 > P \leq 0.05$ (*)	$\|d\| \geq 1.5$ (##)	Suggestive
$P > 0.05$	$P \leq 0.004$ (**)	$\|d\| < 1.0$	No
$P > 0.05$	$P \leq 0.004$ (**)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P > 0.05$	$P \leq 0.004$ (**)	$\|d\| \geq 1.5$ (##)	Significant
$0.004 > P \leq 0.05$ (*)	$P > 0.05$	$\|d\| < 1.0$	No
$0.004 > P \leq 0.05$ (*)	$P > 0.05$	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$0.004 > P \leq 0.05$ (*)	$P > 0.05$	$\|d\| \geq 1.5$ (##)	Suggestive
$0.004 > P \leq 0.05$ (*)	$0.004 > P \leq 0.05$ (*)	$\|d\| < 1.0$	No
$0.004 > P \leq 0.05$ (*)	$0.004 > P \leq 0.05$ (*)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$0.004 > P \leq 0.05$ (*)	$0.004 > P \leq 0.05$ (*)	$\|d\| \geq 1.5$ (##)	Suggestive
$0.004 > P \leq 0.05$ (*)	$P \leq 0.004$ (**)	$\|d\| < 1.0$	No
$0.004 > P \leq 0.05$ (*)	$P \leq 0.004$ (**)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$0.004 > P \leq 0.05$ (*)	$P \leq 0.004$ (**)	$\|d\| \geq 1.5$ (##)	Significant
$P \leq 0.004$ (**)	$P > 0.05$	$\|d\| < 1.0$	No
$P \leq 0.004$ (**)	$P > 0.05$	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P \leq 0.004$ (**)	$P > 0.05$	$\|d\| \geq 1.5$ (##)	Significant
$P \leq 0.004$ (**)	$0.004 > P \leq 0.05$ (*)	$\|d\| < 1.0$	No
$P \leq 0.004$ (**)	$0.004 > P \leq 0.05$ (*)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P \leq 0.004$ (**)	$0.004 > P \leq 0.05$ (*)	$\|d\| \geq 1.5$ (##)	Significant
$P \leq 0.004$ (**)	$P \leq 0.004$ (**)	$\|d\| < 1.0$	No
$P \leq 0.004$ (**)	$P \leq 0.004$ (**)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P \leq 0.004$ (**)	$P \leq 0.004$ (**)	$\|d\| \geq 1.5$ (##)	Significant

[a] The largest Cohen's *d* effect size of all five epochs was used to determine the evidence for a meaningful effect. [b] The symbols in parentheses (*, **, #, ##) are also used in Figures 1, 2, 3, 4, 5 and 6.

2.5.2. Analysis of AUC combined with latencies

Strain differences in AUC combined with latencies were analyzed with analyses of covariance (ANCOVAs) for anxiety, avoidance behavior, risk assessment, arousal, exploration and locomotion. Analogous to the trajectory analyses, separate ANCOVA's compared these values between donor/consomic line and host strain, resulting in 21 analyses for each motivational system/behavioral dimension. Strain was included as main effect, while time of day at test, and season at test were included as covariates. To control for a potentially confounding effect of locomotion or exploration, additional analyses were conducted for anxiety-related behavior which included the integrated z-score for the AUC of locomotion or exploration as a covariate. A bootstrap procedure was applied to the ANCOVA's because the necessary assumptions (homogeneity of variances and/or normal distribution of the residuals) were not always met (Bland et al., 2015). Model-based bootstrapped ANCOVA's (10000 samples) were run using the R-code provided by Mancuso (2020). Main effects of strain and/or the covariates were obtained via the resulting bootstrapped *F*-test and *P*-values. The same threshold was used to determine suggestive ($0.004 \leq P < 0.05$) and significant effects ($P < 0.004$). Effect sizes for the ANCOVA's were reported as partial eta squared (η_p^2) with 95% CI and Cohen's *d* with 95%. The following cut-off limits were used to interpret η_p^2: small effect, $\eta_p^2 \leq 0.03$; medium/moderate effect, $0.03 < \eta_p^2 < 0.10$; large effect, $0.10 \leq \eta_p^2 < 0.20$; very large effect, $\eta_p^2 \geq 0.20$ (Labots et al., 2016). Cohen's *d* values were interpreted following the same guidelines as in the trajectory analyses.

The combination of both effect sizes with the bootstrapped *P* value was again used to determine meaningful strain effects. A very large effect size (Cohen's $d \geq 1.5$ and $\eta_p^2 \geq 0.20$) combined with a significant bootstrap $P < 0.004$ were considered indicative for significant evidence for a chromosome harboring one or more QTLs. A small/medium effect size (Cohen's $d < 1.0$ and/or $\eta_p^2 < 0.10$) combined with $P \geq 0.05$ was considered as no evidence for a QTL harboring chromosome. All other threshold combinations resulted in suggestive evidence. *Large/very large* effect sizes of which the 95% CI contained the value zero were again not included for further evaluation as this amounts to a non-significant ($P \geq 0.05$) result (Nakagawa and Cuthill, 2007).

An overview of the different threshold combinations that were used to establish significant/suggestive evidence for a meaningful difference have been previously published in Labots et al. (2016, Table 3, page 6).

3. Results

3.1. Trajectories

3.1.1. Parental strain analyses

GLMMs compared the behavioral trajectory of each motivational system and behavioral dimension between the donor (A) and the host strain, B6.

Overall anxiety-related behavior differed significantly between the host and donor strain ($\chi^2_{\text{strain (1)}} = 9.78$, $P = 0.0018$; $\chi^2_{\text{strain x epoch (4)}} = 21.61$, $P = 0.0001$, Figure 1-I). *Large* effect sizes and significant *post hoc* comparisons indicated suggestive evidence for higher anxiety in A than in B6 from the 2nd to the last epoch (Figure 1-I, supplementary Table S2). Furthermore, B6 decreased between the first and the last epoch (*large* effect size, suggestive *post hoc* contrast) while anxiety did not differ between the first and the last epoch in A (Figure 1-I, supplementary Table S2). Significant evidence for a meaningful strain difference remained after controlling for a potentially confounding effect of locomotion ($\chi^2_{\text{strain x epoch (4)}} = 22.86$, $P = 0.0001$) or exploration ($\chi^2_{\text{strain (1)}} = 8.08$, suggestive $P = 0.0044$; $\chi^2_{\text{strain x epoch (4)}} = 21.41$, $P = 0.0002$).

Analyses of the separate anxiety-related-dimensions showed that A differed from B6 on avoidance behavior and risk assessment, but not in arousal. The trajectories of avoidance behavior differed between B6 and A ($\chi^2_{\text{strain x epoch (4)}} = 17.22$, $P = 0.0017$, Figure 1-II). *Post hoc* comparisons showed that avoidance behavior decreased between the first and the last epoch in B6, while it remained stable across epochs in A (Figure 1-II, supplementary Table S3). Furthermore, *large* effect sizes and suggestive *post hoc* comparisons showed suggestive evidence for higher avoidance in A than B6 on epochs 3 and 5 (Figure 1-II, supplementary Table S3). A suggestive *P* value for higher avoidance was also found on epoch 4, but this effect was coupled with a medium effect size and therefore did not meet the required combinations of thresholds to be considered meaningful (Figure 1-II, supplementary Table S3).

Overall risk assessment was higher in A than B6, regardless of epoch ($\chi^2_{\text{strain (1)}} = 5.46$, suggestive $P = 0.0195$, Figure 1-III), but *post hoc* comparisons between strains on each epoch showed that risk assessment was higher in A on epoch 4 (suggestive), and not on any of the other epochs (Figure I-III, supplementary Table S4). Furthermore, *post hoc* contrasts showed that risk assessment remained stable between the first and the last epoch in both A and B6 (Figure 1-III, supplementary Table S4). Arousal did not change between the first and the last epoch, regardless of strain, despite a suggestive main effect for epoch ($\chi^2_{\text{epoch (4)}} = 9.89$, $P = 0.0424$, Figure 1-IV). *Post hoc* contrasts assessing the change of arousal between the first and the last epoch for each strain separately confirmed this pattern (Figure 1-IV, supplementary Table S5).

The host and donor strain also differed in activity behavior, although the directionality of the trajectories was not very different between strains. Exploration increased regardless of strain ($\chi^2_{\text{epoch (4)}} = 58.82$, $P = 0.0000$) but overall exploration was lower in A compared to B6 ($\chi^2_{\text{strain (1)}} = 34.10$, $P = 0.0000$, Figure 1-V). *Post hoc* comparisons between strains on each epoch indicated suggestively higher exploration on epoch 2 in B6 than in A, but this effect was not considered meaningful as it did not meet the required combination of thresholds (Figure 1-V, supplementary Table S6). *Post hoc* contrasts between the first and the last epoch for each strain separately showed that exploration increased in both A and B6 (Figure 1-V, supplementary Table S5).

Locomotion differed between strains ($\chi^2_{\text{strain (1)}} = 244.795$, $P = 0.0001$; $\chi^2_{\text{strain x epoch (4)}} = 15.15$, suggestive $P = 0.0044$, Figure 1-VI). *Post hoc* comparisons showed that locomotion remained stable between the first and the last epoch in both strains (Figure I-VI, supplementary Table S7). Most notably, *very large* effect sizes combined with significant *post hoc* comparisons locomotion showed that locomotion was significantly higher in B6 than A on all epochs (Figure 1-VI, supplementary Table S7).

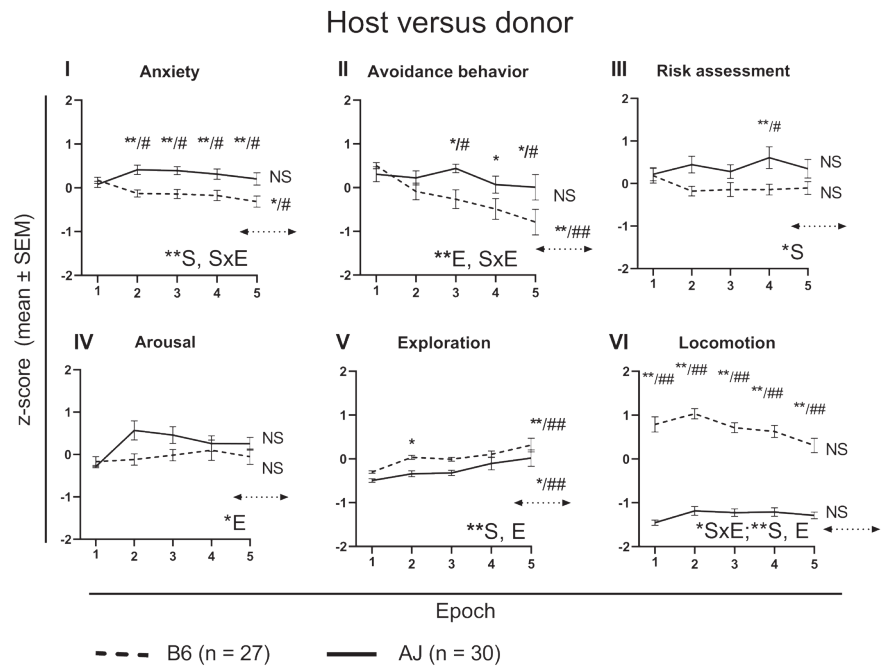


Figure 1. Trajectories of the host (B6) and donor (A) strain for I) anxiety, II) avoidance behavior, III) risk assessment, IV) arousal, V) exploration and VI) locomotion. Results expressed as integrated behavioral z-scores (anxiety, avoidance behavior, arousal, exploration) and behavioral z-score (risk assessment, locomotion). All results presented as means with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = $0.004 \leq P < 0.05$ and significant at ** = $P < 0.004$. S indicates an effect of strain; E indicates an effect of epoch and S x E indicates an interaction between strain and epoch. *Post hoc* comparisons between B6 and A on each epoch: suggestive difference at * = $0.00200 \leq P < 0.025321$, significant difference at ** = $P < 0.00200$ for epochs 1 and 5. Suggestive difference at * = $0.004 \leq P < 0.05$ and significant difference at ** = $P < 0.004$ for epochs 2, 3 and 4. Dashed arrow: *Post hoc* contrast between epochs 1 and 5 (assessing the change of anxiety across time) specified for A and B6 separately: suggestive difference at * = $0.00200 \leq P < 0.025321$, significant difference at ** = $P < 0.00200$. The Cohen's *d* effect size shows the relative magnitude of each difference between strains, or between epochs (*post hoc* only). Effect sizes are indicated with # = *Large* ($1.0 \leq |d| < 1.5$); ## = *Very large* ($|d| \geq 1.5$).

3.1.2. Consomic strain survey

The survey comparing the trajectories between the panel of consomic lines and B6 identified two lines with significant and/or suggestive evidence for a meaningful QTL for anxiety-related behavior: CSS-10 and CSS-19.

The trajectories for overall anxiety only differed from B6 in CSS-19 ($\chi^2_{\text{strain (1)}} = 9.46$, $P = 0.0021$). Suggestive/significant *post hoc* comparisons combined with *very large* effect sizes indicated suggestive evidence for higher anxiety in CSS-19 compared to B6 on epochs 2 and 5, and significant evidence for higher anxiety on epoch 4 (Figure 2, supplementary Table S2). This strain effect remained when controlling for a potential confounding effect of locomotor activity, by incorporating the z-score for locomotion in the model ($\chi^2_{\text{strain (1)}} = 9.59$, $P = 0.0019$). This effect also remained after controlling for exploration ($\chi^2_{\text{strain (1)}} = 7.54$, suggestive $P = 0.0060$). Significant/suggestive effects of epoch showed that anxiety decreased regardless of strain in all consomic versus host comparisons, with the exception of CSS-11 and CSS-13, in which overall anxiety did not change over time (Figure 2, supplementary Table S2).

The trajectory of avoidance behavior in CSS-19 also differed from B6 ($\chi^2_{\text{strain (1)}} = 8.49$, $P = 0.0036$; $\chi^2_{\text{strain x epoch (4)}} = 12.97$, suggestive $P = 0.0114$). While B6 decreased, elevated levels of avoidance behavior remained stable in CSS-19 between epoch 1 and epoch 5 (Figure 3, supplementary Table S3). *Large* effect sizes and suggestive/significant *post hoc* comparisons indicated suggestive evidence for higher avoidance behavior in CSS-19 than in B6 on epochs 2, 3, 4 and 5 (Figure 3, supplementary Table S3). The remaining consomic lines all decreased avoidance behavior in concordance with B6, as indicated by significant main effects of epoch on each comparison between host and consomic line (Figure 3, supplementary Table S3).

The survey also revealed suggestive evidence of overall higher risk assessment in CSS-19 ($\chi^2_{\text{strain (1)}} = 7.46$, suggestive $P = 0.0063$, Figure 4). Further *post hoc* inspection of strain differences on each epoch showed significant evidence for higher risk assessment in CSS-19 on epoch 2, and suggestive evidence for higher risk assessment on epoch 5 (Figure 4, supplementary Table S4). In addition, risk assessment trajectories differed between CSS-10 and B6 ($\chi^2_{\text{strain x epoch (4)}} = 33.45$, $P = 0.0000$), see Figure 4. Risk assessment was significantly higher in CSS-10 than B6 in the first two epochs as indicated by *very large* effect sizes and significant *post hoc* comparisons (Figure 4, supplementary Table S4), and decreased between the first and last epoch (Figure 4, supplementary Table S4). In the remaining consomic lines risk assessment did not differ from B6 (Figure 4).

Lastly, the change of arousal over time did not differ from B6 in any of the consomic lines (Figure 5).

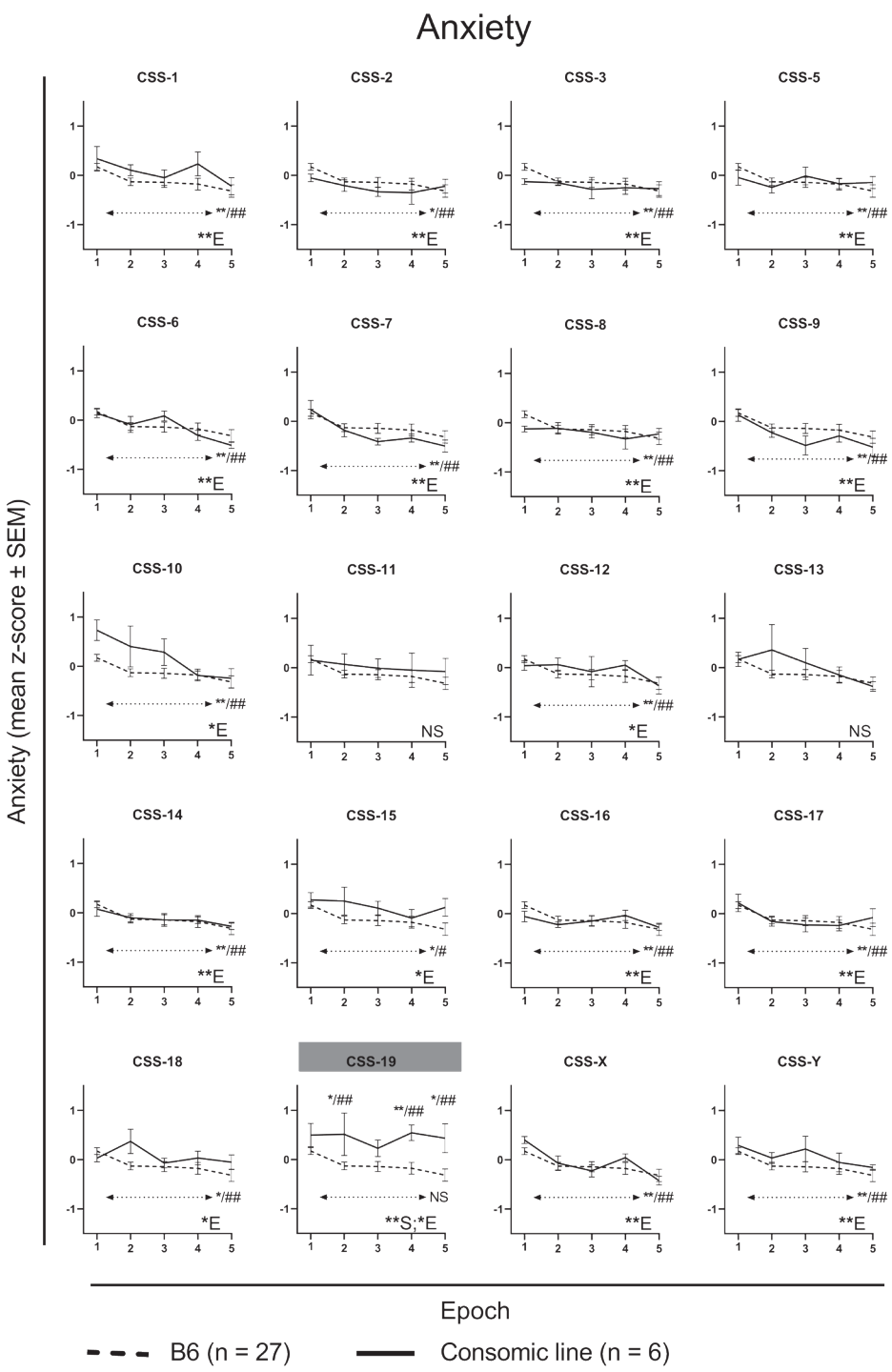


Figure 2. Consomic strain survey on anxiety trajectories in a CSS-panel, comparing the host strain (dashed line) and each CSS-line (black line) in a separate panel. Results expressed as mean integrated behavioral z-scores and presented as means with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = 0.004 ≤ P < 0.05 and significant at ** = P < 0.004. S indicates an effect of strain; E indicates an effect of epoch. *Post hoc* comparisons between B6 and CSS on each epoch (in case of a main effect of strain): suggestive difference at * = 0.00200 ≤ P < 0.025321, significant difference at ** = P < 0.00200 for epochs 1 and 5. Suggestive difference at * = 0.004 ≤ P < 0.05 and significant difference at ** = P < 0.004 for epochs 2, 3 and 4. Dashed arrow: *Post hoc* contrast of anxiety between epochs 1 and 5 (assessing the change of anxiety across time) for both strains combined (in case of E): suggestive difference at * = 0.004 ≤ P < 0.05, significant difference at ** = P < 0.004. The Cohen's d effect size shows the relative magnitude of each difference between strains, or between epochs (*post hoc* only). Effect sizes are indicated with # = *Large* (1.0 ≤ |d| < 1.5); ## = *Very large* (|d|≥ 1.5). Grey panel: significant evidence for a meaningful QTL for that consomic line.

Next to anxiety-related behavior, we assessed differences in activity behavior, exploration and locomotion. Overall exploration had been lower in A compared to B6. Suggestive/significant interactions between strain and epoch and/or strain effects indicated that exploration trajectories differed from B6 in the majority (80%) of consomic lines: CSS-1, CSS-2, CSS-3, CSS-6, CSS-7, CSS-8, CSS-9, CSS-10, CSS-12, CSS-13, CSS-15, CSS-16, CSS-17, CSS-18, CSS-19, CSS-X (Figure 6, Supplementary Table S1). *Very large* effect sizes and significant *post hoc* comparisons between the first and the last epoch revealed a significant increase in exploration in all these lines (Figure 6, supplementary Table S6). In CSS-5, CSS-11, CSS-14 and CSS-Y a significant increase was observed between the first and the last epoch, regardless of strain (Figure 6, supplementary Table S6). *Post hoc* comparisons showed that the found interactions were predominantly driven by suggestively lower exploration than B6 on epoch 1 (CSS-3, CSS-7, CSS-9, CSS-16) or significantly/suggestively lower exploration on epochs 1 and 2 (CSS-1, CSS-6, CSS-10, CSS-12, CSS-13, CSS-15, CSS-17, CSS-18, CSS-19, CSS-X), Figure 6, supplementary Table S6. In these lines, exploration increased up to similar (i.e., not suggestive/significantly different) levels as B6 in the remaining epochs (Figure 6, supplementary Table S6). In CSS-2, *large/very large* effect sizes and a suggestive *post hoc* comparison between B6 on each epoch showed suggestive evidence for lower exploration in CSS-2 on epochs 1 and 3 (Figure 6, supplementary Table S6).

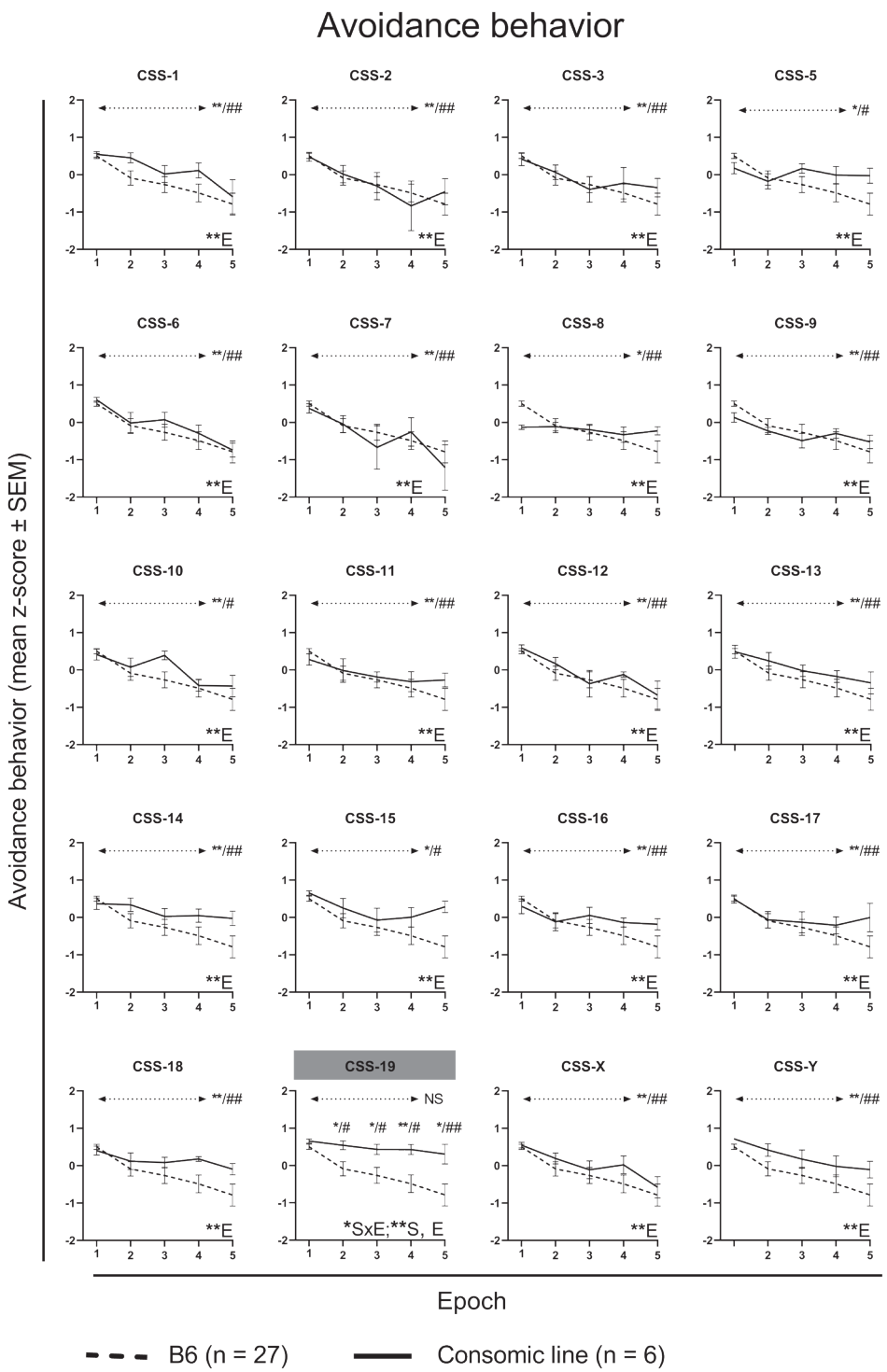


Figure 3. Consomic strain survey on avoidance behavior trajectories in a CSS-panel, comparing the host strain (dashed line) and each CSS-line (black line) in a separate panel. Results expressed as integrated behavioral z-score and presented as mean with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = 0.004 ≤ *P* < 0.05 and significant at ** = *P* < 0.004. S indicates an effect of strain; E indicates an effect of epoch and S x E indicates an interaction between strain and epoch. *Post hoc* comparisons between B6 and CSS on each epoch (in case of a S, or S x E): suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200 for epochs 1 and 5. Suggestive difference at * = 0.004 ≤ *P* < 0.05 and significant difference at ** = *P* < 0.004 for epochs 2, 3 and 4. Dashed arrow: *Post hoc* contrast of avoidance behavior between epochs 1 and 5 (assessing the change of avoidance behavior across time) for that particular CSS (in case of S and/or S x E) or for both strains combined (in case of E). In case of S/S x E: Suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200. In case of E: Suggestive difference at * = 0.004 ≤ *P* < 0.05, significant difference at ** = *P* < 0.004. The Cohen's *d* effect size shows the relative magnitude of each difference between strains, or between epochs (*post hoc* only). Effect sizes are indicated with # = *Large* (1.0 ≤ |*d*| < 1.5); ## = *Very large* (|*d*|≥ 1.5). Grey panel: significant evidence for a meaningful QTL for that consomic line.

For locomotion, the survey revealed significant/suggestive evidence for a meaningful QTL for locomotor activity in all consomic lines, in the form of significant or suggestive effects for strain and/or an interaction between strain and epoch (Figure 7, supplementary Table S1). As described above locomotor activity was markedly lower in the donor strain A than in B6 on all 5 epochs. The consomic strain survey indeed confirmed significant lower levels of locomotion on the first epoch in all consomic lines (Figure 7, supplementary Table S7) but *post hoc* contrasts showed an increase in locomotion between the first and the last epoch in the majority of consomic lines, with suggestive evidence (*large* effect sizes combined with a significant *post hoc* test) for an increase in CSS-2, CSS-9, CSS-12, CSS-13, CSS-14, and a significant (*post hoc* contrast combined with *very large* effect sizes) increase in CSS-1, CSS-7, CSS-15, CSS-16, CSS-17, CSS-18, CSS-19, CSS-X and CSS-Y (Figure 7, supplementary Table S7). Locomotion did not differ significantly/suggestively between the first and the last epoch in CSS-3, CSS-5, CSS-6, CSS-8, CSS-10, and CSS-11 (Figure 7, supplementary Table S7).

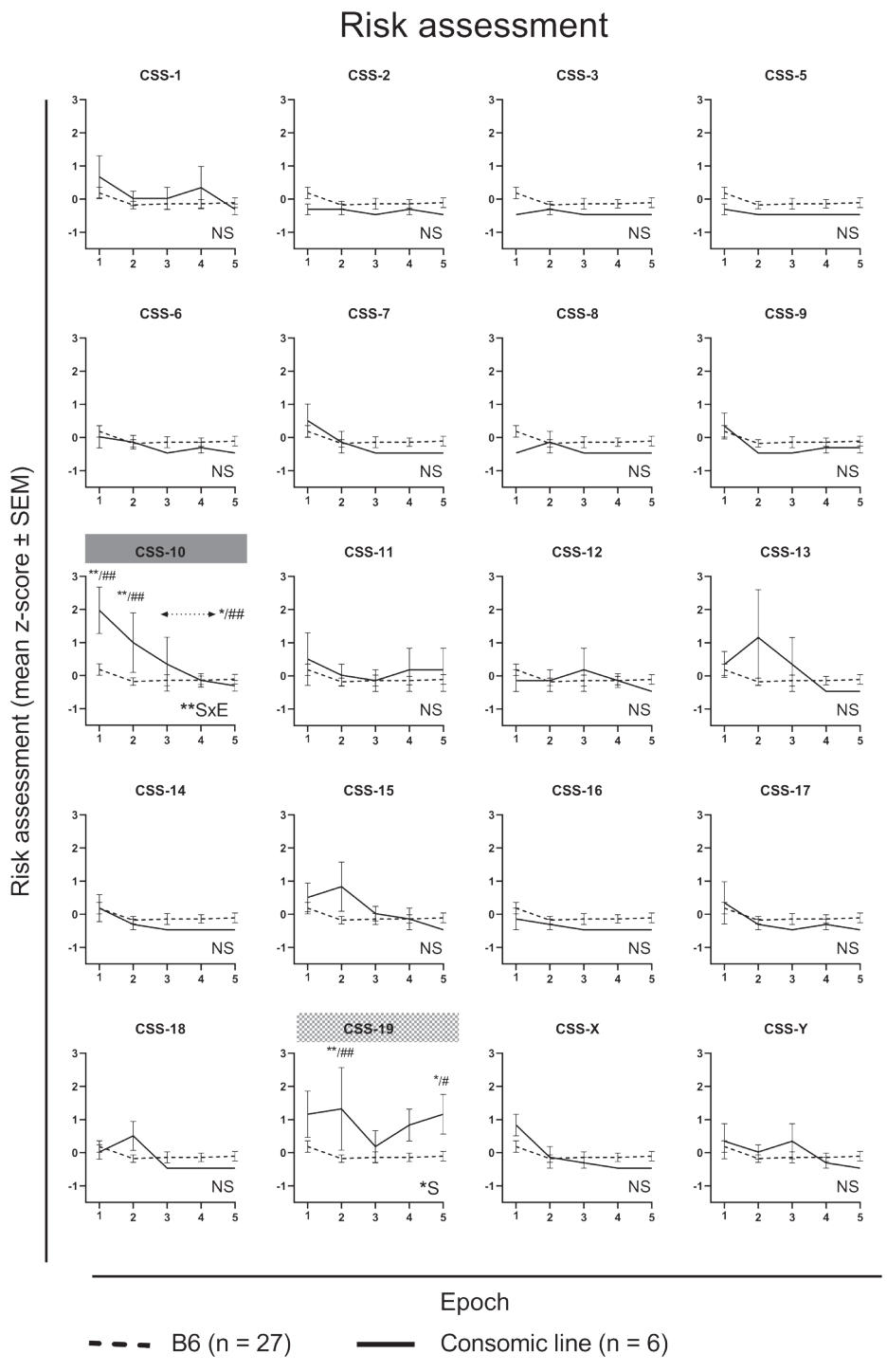


Figure 4. Consomic strain survey on risk assessment trajectories in a CSS-panel, comparing the host strain (dashed line) and each CSS-line (black line) in a separate panel. Results expressed as behavioral z-score and presented as mean with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = 0.004 ≤ *P* < 0.05 and significant at ** = *P* < 0.004. S indicates an effect of strain; S x E indicates a found interaction between strain and epoch. *Post hoc* comparisons between B6 and CSS on each epoch (in case of S and/or S x E): suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200 for epochs 1 and 5. Suggestive difference at * = 0.004 ≤ *P* < 0.05 and significant difference at ** = *P* < 0.004 for epochs 2, 3 and 4. Dashed arrow: *Post hoc* contrast of risk assessment between epochs 1 and 5 (assessing the change of risk assessment across time) for that particular CSS (in case of S and/or S x E): suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200. The Cohen's *d* effect size shows the relative magnitude of each difference between strains, or between epochs (*post hoc* only). Effect sizes are indicated with # = *Large* (1.0 ≤ |*d*| < 1.5); ## = *Very large* (|*d*|≥ 1.5). Grey panel: significant evidence; checkered panel: suggestive evidence for a meaningful QTL for that consomic line.

In some lines, locomotion only differed significantly/suggestively on epoch 1, after which it increased to similar (i.e., not suggestive/significantly different) levels of locomotion as B6 on epochs 2 to 5: CSS-2, CSS-5, CSS-7, CSS-8, CSS-9, CSS-14, CSS-17, CSS-19, CSS-Y (Figure 7, supplementary Table S7). In other lines, the markedly lower levels of locomotion persisted up to epoch 2 (CSS-1, CSS-12, CSS-13, CSS-X), epoch 3 (CSS-6, CSS-15) or epoch 4 (CSS-10, CSS-11, CSS-18), after which they increased to similar levels as the host strain (Figure 7, supplementary Table S7). In CSS-3 locomotion was lower on all five epochs. Finally, in CSS-16 locomotion was lower on epochs 1, 2 and 4 (Figure 7, supplementary Table S7).

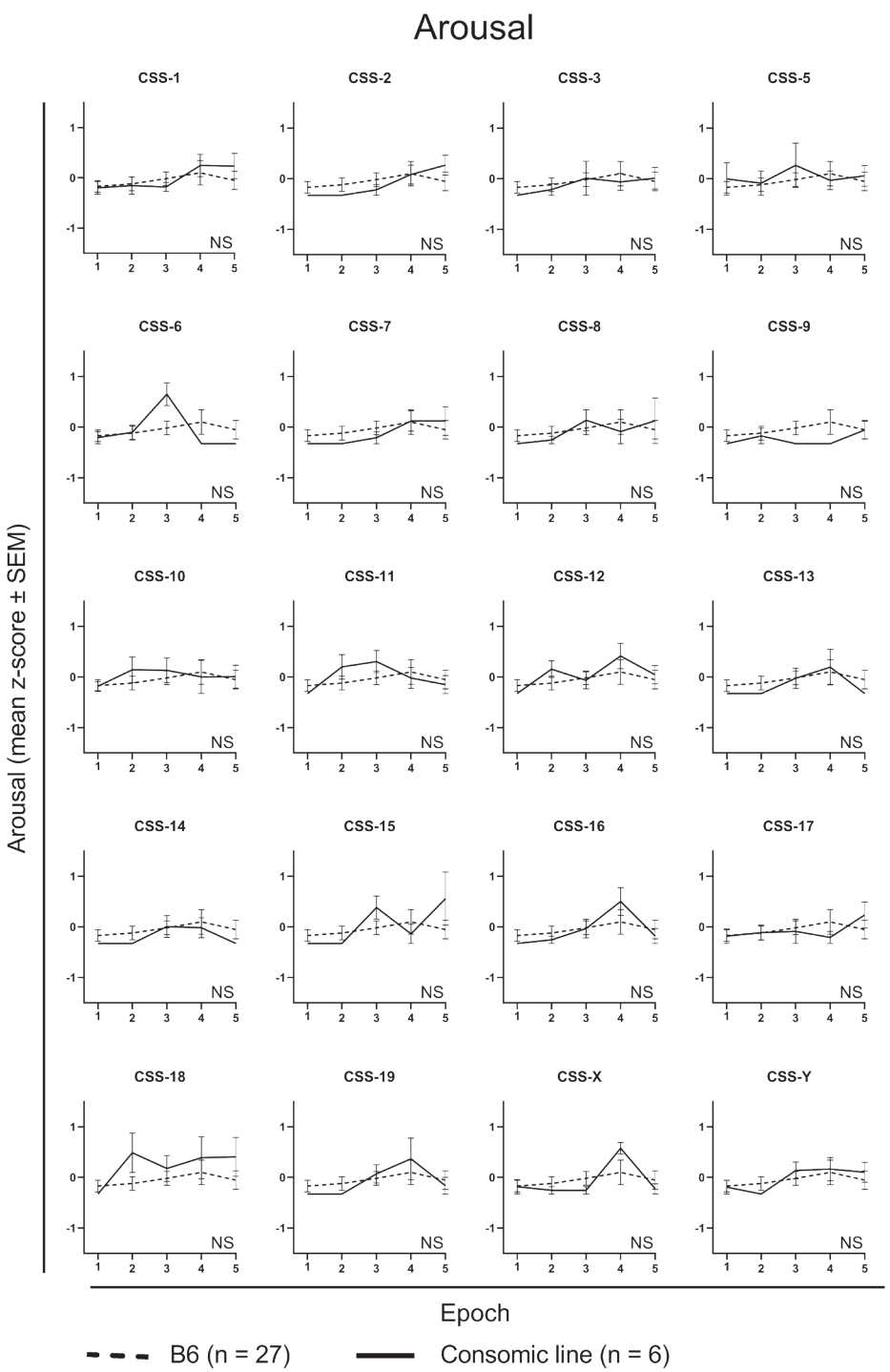


Figure 5. Consomic strain survey on arousal trajectories in a CSS-panel. Each panel compares the host strain (dashed line) and one CSS-line (black line). Results expressed as integrated behavioral z-score and presented as mean with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = $0.004 \leq P < 0.05$ and significant at ** = $P < 0.004$. NS indicates no significant difference between B6 and that consomic line.

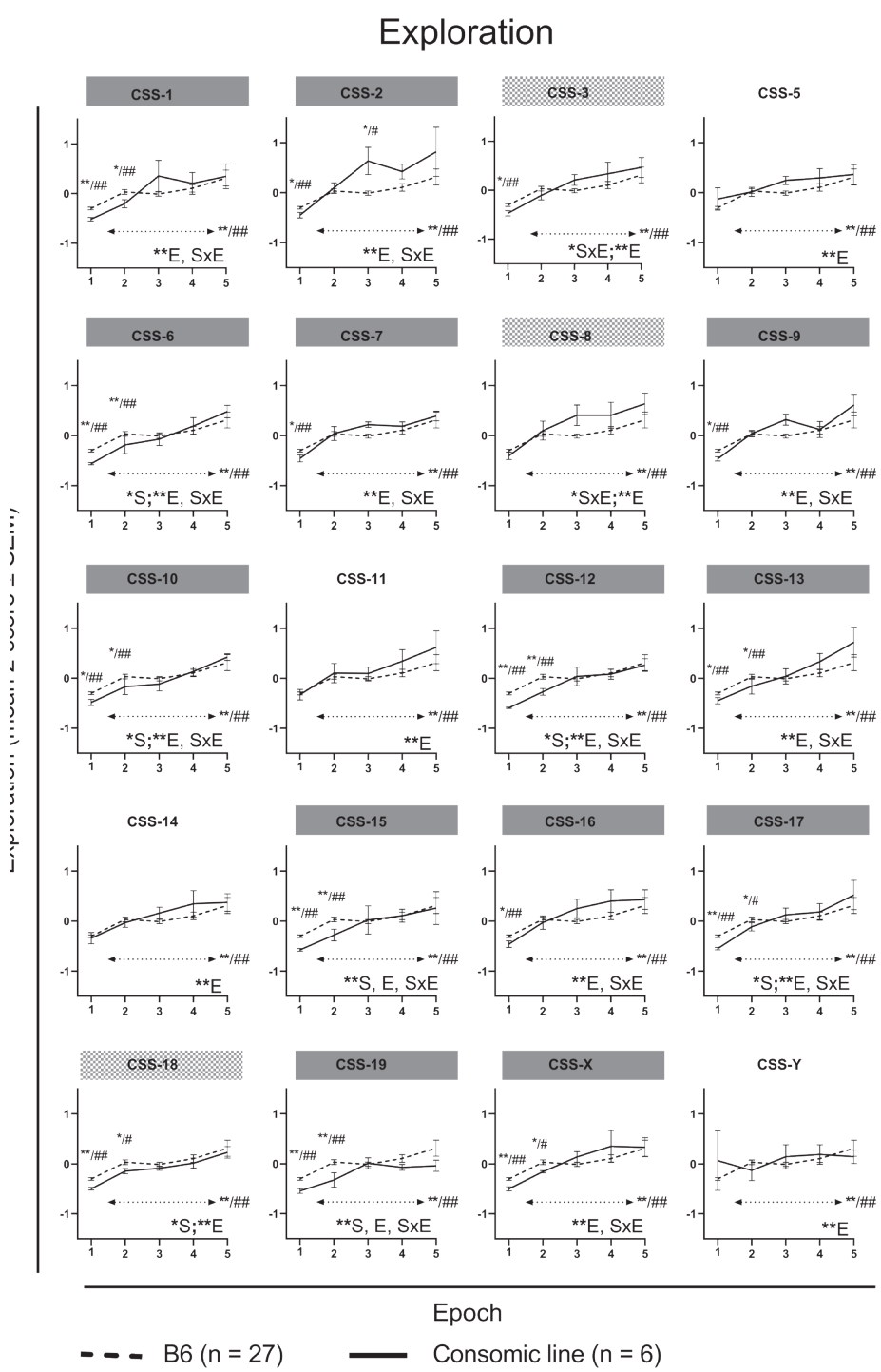


Figure 6. Consomic strain survey on exploration trajectories in a CSS-panel, comparing the host strain (dashed line) and each CSS-line (black line) in a separate panel. Results expressed as integrated behavioral z-score and presented as mean with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = 0.004 ≤ *P* < 0.05 and significant at ** = *P* < 0.004. S indicates an effect of strain; E indicates an effect of epoch and S x E indicates an interaction between strain and epoch. *Post hoc* comparisons between B6 and CSS on each epoch (in case of S and/or S x E): suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200 for epochs 1 and 5. Suggestive difference at * = 0.004 ≤ *P* < 0.05 and significant difference at ** = *P* < 0.004 for epochs 2, 3 and 4. Dashed arrow: *Post hoc* contrast of exploration between epochs 1 and 5 (assessing the change of exploration across time) for that particular CSS (in case of S and/or S x E) or for both strains combined (in case of E). In case of S/S x E: Suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200. In case of E: Suggestive difference at * = 0.004 ≤ *P* < 0.05, significant difference at ** = *P* < 0.004. The Cohen's *d* effect size shows the relative magnitude of each difference between strains, or between epochs (*post hoc* only). Effect sizes are indicated with # = *Large* (1.0 ≤ |*d*| < 1.5); ## = *Very large* (|*d*|≥ 1.5). Grey panel: significant evidence; checkered panel: suggestive evidence for a meaningful QTL for that consomic line.

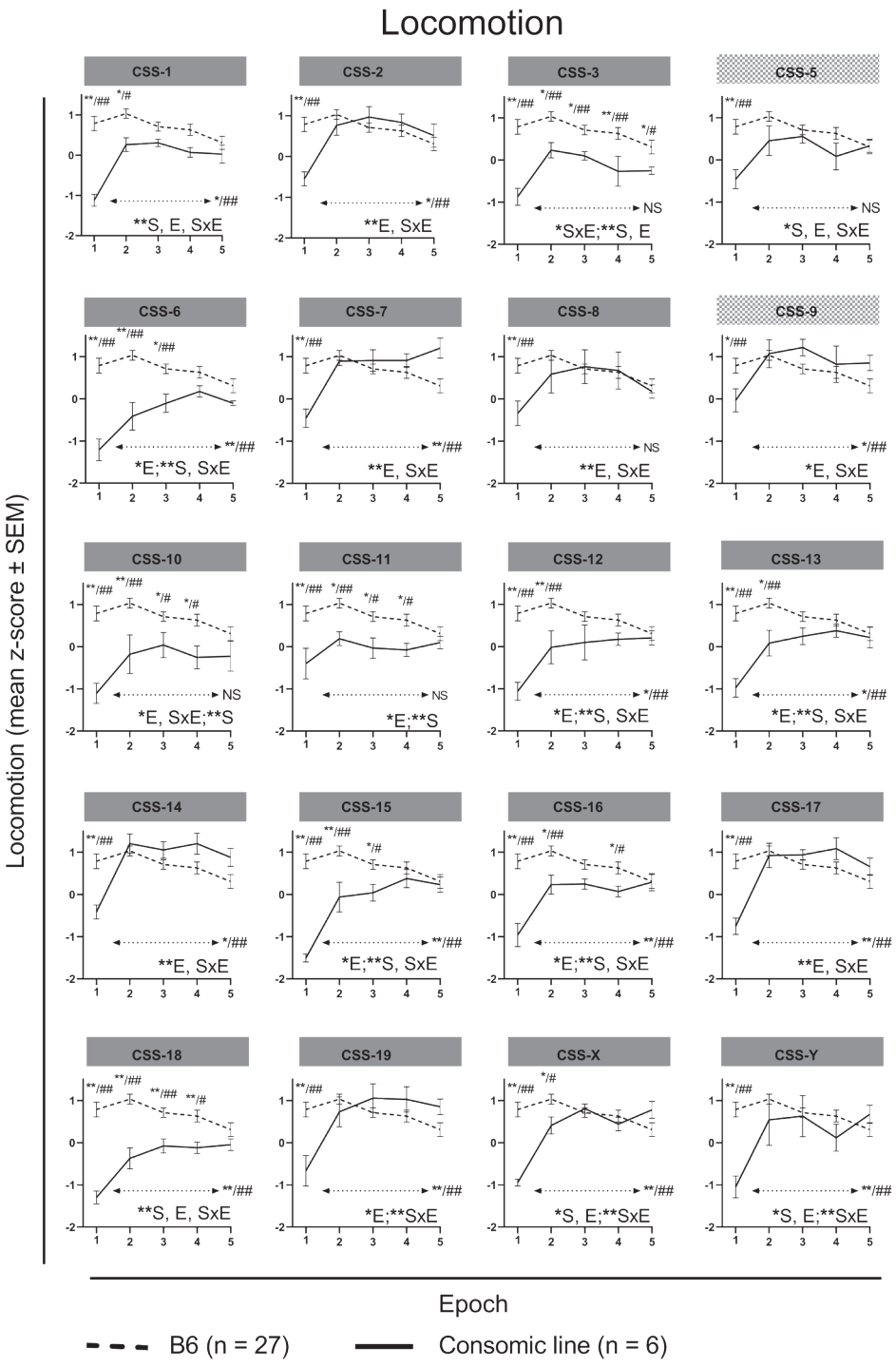


Figure 7. Consomic strain survey on locomotion trajectories in a CSS-panel, comparing the host strain (dashed line) and each CSS-line (black line) in a separate panel. Results expressed as behavioral z-score and presented as mean with SEM. Main and/or interaction effects resulting from GLMMs were suggestive at * = 0.004 ≤ *P* < 0.05 and significant at ** = *P* < 0.004. S indicates an effect of strain; E indicates an effect of epoch and S x E indicates an interaction between strain and epoch. *Post hoc* comparisons between B6 and CSS on each epoch (in case of S and/or S x E): suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200 for epochs 1 and 5. Suggestive difference at * = 0.004 ≤ *P* < 0.05 and significant difference at ** = *P* < 0.004 for epochs 2, 3 and 4. Dashed arrow: *Post hoc* contrast of locomotion between epochs 1 and 5 (assessing the change of locomotion across time) for that particular CSS (in case of S and/or S x E) or for both strains combined (in case of E). In case of S/S x E: Suggestive difference at * = 0.00200 ≤ *P* < 0.025321, significant difference at ** = *P* < 0.00200. In case of E: Suggestive difference at * = 0.004 ≤ *P* < 0.05, significant difference at ** = *P* < 0.004. The Cohen's *d* effect size shows the relative magnitude of each difference between strains, or between epochs (*post hoc* only). Effect sizes are indicated with # = *Large* (1.0 ≤ |*d*| < 1.5); ## = *Very large* (|*d*|≥ 1.5). Grey panel: significant evidence; checkered panel: suggestive evidence for a meaningful QTL for that consomic line.

3.2. AUC combined with latencies

3.2.1. Parental strain analyses

Comparing the AUC combined with latency between A and B6 resulted in a suggestive meaningful difference in overall anxiety (*large/very large* effect sizes, $d = 1.290$, $\eta_p^2 = 0.296$, $F_{(1,53)} = 20.28$, bootstrap $P = 0.0001$, Figure 8-I). This effect did not reach the criteria for a suggestive meaningful difference after controlling for locomotion, as indicated by *medium/very large* effect sizes and a suggestive *P*-value ($d = 0.85$, $\eta_p^2 = 0.31$, $F_{(1,53)} = 4.82$, bootstrap $P = 0.0323$). The meaningful difference in anxiety remained however after controlling for exploration (*large/very large* effect sizes, $d = 1.420$, $\eta_p^2 = 0.390$, $F_{(1,53)} = 24.02$, bootstrap $P = 0.0001$).

ANCOVA's comparing the host and donor strain on separate anxiety-related dimensions revealed no meaningful difference in avoidance behavior (*small/medium* effect size, $d = 0.49$, $\eta_p^2 = 0.066$, $F_{(1,53)} = 2.97$, bootstrap $P = 0.0880$, Figure 8-II). Risk assessment differed significantly between strains (Figure 8-IV, $F_{(1,53)} = 10.35$, bootstrap $P = 0.0027$), but this difference did not meet the required combination of thresholds for being meaningful due to a *medium/large* effect size ($d = 0.925$, $\eta_p^2 = 0.168$, supplementary Table S9). Finally, there was a suggestive meaningful difference in arousal between the parental strains (*large/very large* effect size, $d = 1.180$, $\eta_p^2 = 0.247$, $F_{(1,53)} = 16.72$, bootstrap $P = 0.0002$, Figure 8-III).

Locomotion differed significantly and meaningful between the parental strains (*very large* effect size, d = -1.91, η_p^2 = 0.487, $F_{(1,53)}$ = 44.28, bootstrap P = 0.0001, Figure 9-II), while exploration did not differ between strains (*small* effect size, d = -0.22, η_p^2 = 0.024, $F_{(1,53)}$ = 0.57, bootstrap P = 0.4543, Figure 9-I).

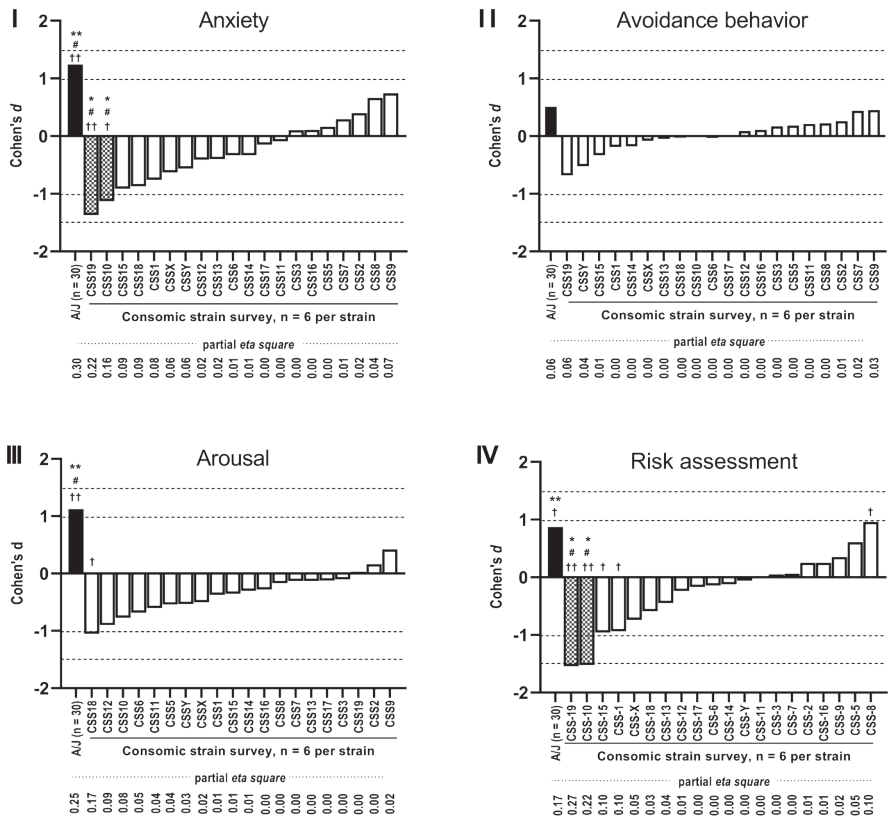


Figure 8. Effect sizes for the AUC combined with latencies (integrated behavioral z-scores) of (I) overall anxiety-related behavior, (II) avoidance behavior, (III) arousal and (IV) risk assessment. The effect sizes (Cohen's d and η_p^2) show the relative magnitude of strain differences compared to the host strain, B6. Suggestive evidence resulting from ANCOVAs indicated for the donor strain (A)/consomic lines with * = 0.004 ≤ P < 0.05. All other significant tests indicated with ** = P < 0.004. Chromosomal effect sizes are indicated with # = *Large* (1.0 ≤ |d| < 1.5); ## = *Very large* (|d|≥ 1.5), † = *Large* (0.10 ≤ η_p^2 < 0.20) and †† = *Very large* (η_p^2 ≥ 0.20). Black bar: donor strain (A), checkered bars: CSS with suggestive evidence for a meaningful QTL, white bars: CSS with no evidence for a meaningful QTL.

3.2.2. Consomic strain survey

The consomic strain survey of the AUC combined with latencies regarding overall anxiety resulted in suggestive evidence for a meaningful QTL for CSS-10 and CSS-19 (CSS-10, *large* effect size, d = -1.35, η_p^2 = 0.156, $F_{(1,29)}$ = 4.86, suggestive bootstrap P = 0.03370; CSS-19, *large/very large* effect size, d = -1.46, η_p^2 = 0.216; $F_{(1,29)}$ = 6.47, suggestive bootstrap P = 0.0168; Figure 8-I). This suggestive evidence for a meaningful QTL remained for CSS-19 when controlling for locomotor activity (*large/very large* effect size, d = -1.42, η_p^2 = 0.230, $F_{(1,29)}$ = 6.04, bootstrap P = 0.0190), but not when controlling for exploration (*medium/very large* effect size, d = -0.856, η_p^2 = 0.360, $F_{(1,29)}$ = 1.93, bootstrap P = 0.1796). For CSS-10, the suggestive evidence for a meaningful difference in anxiety disappeared after controlling for locomotion (*medium/large* effect size, d = -0.683, effect size, η_p^2 = 0.160, $F_{(1,29)}$ = 0.76, bootstrap P = 0.3856) and exploration (*medium/very large* effect size, d = -0.736, η_p^2 = 0.230, $F_{(1,29)}$ = 1.27, bootstrap P = 0.2710). No confounding effect of locomotion on overall anxiety-related behavior was found in any of the other consomic lines (ANCOVA output available from the corresponding author upon request).

Analyzing the separate anxiety-related dimensions also revealed suggestive evidence for a meaningful QTL for risk assessment in CSS-10 and CSS-19 (CSS-10, *large/very large* effect size, d = -1.48, η_p^2 = 0.221, $F_{(1,29)}$ = 4.88, bootstrap P = 0.0364; CSS-19, *very large* effect size, d = -1.54, η_p^2 = 0.266, $F_{(1,293)}$ = 6.62, bootstrap P = 0.0158), see Figure 8-IV. Comparing risk assessment between the consomic lines and B6 furthermore resulted in *large* η_p^2 effect sizes for CSS-1, CSS-8 and CSS-15, but these effects did not meet the required combination of thresholds for a meaningful QTL (CSS-1, bootstrap P = 0.1476; bootstrap P = 0.1528; bootstrap P = 0.1668, Figure 8-IV, supplementary Table S9). No suggestive/significant evidence for a meaningful QTL was found for avoidance behavior and arousal, in any of the lines (Figure 8-II, 8-III, supplementary Table S8). For arousal, a *large* η_p^2 effect size was found for CSS-18, but the corresponding P-value did not exceed the required threshold (bootstrap P = 0.1337; Figure 8-III, supplementary Table S9).

For exploration, no suggestive/significant evidence for a meaningful QTL was found in any of the consomic lines (Figure 9-I, supplementary Table S8). *Large* effect sizes and/or a suggestive P-value were found for CSS-2, CSS-5, CSS8 and CSS-19 but in none of these lines the effects met the required combination

of thresholds (Figure 9-I; CSS-2, bootstrap P = 0.4108; CSS-5, bootstrap P = 0.3402; CSS-8, bootstrap P = 0.3442; CSS-19, η_p^2 = 0.0053, bootstrap P = 0.0498, supplementary Table S10).

With respect to locomotion, the survey resulted in significant evidence for meaningful QTL(s) in a large number of lines: CSS-1, CSS-3, CSS-6, CSS-10, CSS-12, CSS-15, CSS-16, CSS-18 (Figure 9-II, supplementary Tables S8, S10). Suggestive evidence for a meaningful QTL was found in an additional number of consomic lines: CSS-11, CSS-13 and CSS-X (Figure 9-II, supplementary Tables S8, S10). *Large/very large* effect sizes were also found for CSS-5 and CSS-Y, but these effects were not combined with a suggestive/significant P-value (Figure 9-II; CSS-5, bootstrap P = 0.1035; CSS-Y, bootstrap P = 0.6796, supplementary Table S10).

4. Discussion

In this survey we asked whether the genetic regions modulating anxiety-related behavior in A and B6 in the mHB also affect the change of anxiety-related behavior over time, i.e. the extent to which these strains are able to adapt to a novel environment. We used a multidimensional approach to assess the quality of anxiety responses, and used effect size measures in combination with statistical significance as a selection method of consomic lines, as was recommended by Labots et al., (2016). Previous analyses on part of this dataset found that the parental strains differed in anxiety-related and activity behavior in the mHB: A mice were characterized by high anxiety and low activity, while B6 displayed a highly active and low anxious profile (Laarakker et al., 2008; Labots et al., 2016). When analyzing the change of anxiety-related and activity behavior over time we found the same contrasting profiles: anxiety-related behavior was higher in A than in B6, while B6 were more active than A mice.

In B6, trajectories of overall anxiety-related behavior and avoidance behavior decreased over the course of five minutes. In A, high levels of anxiety-related behavior and avoidance behavior remained elevated throughout the trial. Overall risk assessment was higher in A than B6 regardless of epoch, while arousal did not differ between strains (Fig. 1). These differential patterns were in line with the analysis of the AUC combined with latencies, where a positive effect size in A compared to B6 indicated a less pronounced change in anxiety-related behavior, avoidance behavior and risk assessment over time (Fig. 8).

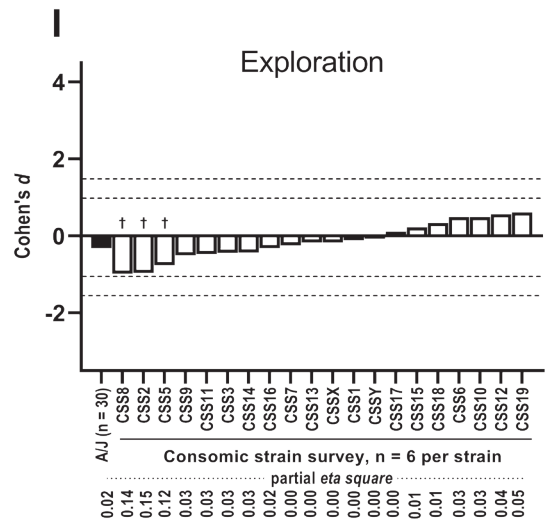


Figure 9. Effect sizes for the AUC combined with latencies (integrated behavioral z-scores) of (I) exploration, (II) locomotion. The effect sizes (Cohen's d and η_p^2) show the relative magnitude of strain differences compared to the host strain, B6. Suggestive evidence resulting from ANCOVAs indicated for the donor strain (A)/consomic lines with * = 0.004 ≤ P < 0.05. All other significant tests indicated with ** = P < 0.004. Chromosomal effect sizes are indicated with # = *Large* (1.0 ≤ |d| < 1.5); ## = *Very large* (|d|≥ 1.5), † = *Large* (0.10 ≤ η_p^2 < 0.20) and †† = *Very large* (η_p^2 ≥ 0.20). Black bar: donor strain (A), grey bars: CSS with significant evidence for a meaningful QTL, checkered bars: CSS with suggestive evidence for a meaningful QTL, white bars: CSS with no evidence for a meaningful QTL.

With respect to activity, significantly lower levels of locomotion remained stable over the course of the trial in A, while high levels of locomotion decreased in B6. In addition, overall exploratory activity was higher in B6 and increased regardless of strain (Fig. 1). These differences in locomotor activity warranted a further investigation of a potentially confounding effect of locomotor activity on the expression of overall anxiety-related behavior (Ohl, 2003). Controlling for this variable however, by including it as a covariate when analyzing anxiety trajectories, did not reveal a clear confounding effect of this behavior on anxiety-related behavior. The persistent high levels of anxiety-related behavior in A thus indeed appear indicative of high anxiety and not the mere result of low activity levels.

The observed temporal profile of B6 corresponds with previous literature in which the change of behavior over time was characterized by both a decrease in anxiety-related and activity behavior (Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Bolivar et al., 2000; Bothe et al., 2005). High levels of avoidance behavior that remained stable across time have indeed been previously reported in A (Gershenfeld and Paul, 1997; Bolivar et al., 2000), but see Bothe et al., (2005) for a report of increased avoidance behavior across time. Likewise, stable low levels of locomotor activity have also been reported previously (Logue et al., 1997; Bothe et al., 2005), but at the same time have been observed to increase (Bolivar et al., 2000; Bolivar, 2009) or decrease over time (Gershenfeld et al., 1997) in A. These incongruent results may be partly related to procedural differences between these studies. For one, some studies measured the change of behavior within the same trial (present study; Gershenfeld and Paul, 1997; Logue et al., 1997; Bolivar, 2009), while others measured the change of behavior across test sessions (Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Bolivar et al., 2000; Bothe et al., 2005; Bolivar, 2009).

The majority of this research was conducted in the field of behavioral habituation of locomotor activity, which emphasizes the change of locomotor behavior during repeated or prolonged exposure to a novel environment/ stimulus (e.g. Bolivar, 2000; Bothe et al., 2005; Bolivar 2009). Within this field of research it is common to distinguish between intrasession and intersession habituation (Bolivar, 2009): the change of activity within the same test session (intra), or across test sessions (inter). These two phenomena are believed to represent different constructs, where the first measures adaptive capacities to a novel environment, while the latter reflects both adaptivity and memory of this environment (Bolivar, 2009). It is currently unknown whether a similar

distinction applies to habituation of anxiety-related behavior measured within the same session, or measured across sessions. Measuring the change of behavior within the same trial, or across sessions however has been observed to produce different effects. In A mice for example, locomotor activity has been reported to remain stable when assessed in a five minute trial (present study, Fig. 1; Logue et al., 1997), decrease over the course of a 15-minute open field trial (Bolivar, 2009) and increase when assessed over three consecutive days (Bolivar et al., 2000).

The above example also demonstrates how another factor may affect the temporal patterns of behavior: trial length. This factor indeed differed between five minutes (present study; Logue et al., 1997) and fifteen minutes (Gershenfeld and Paul, 1997; Bolivar, 2009) in the studies that measured change of behavior within the same test session. Bolivar (2009) compared habituation of locomotor activity in several mouse inbred strains and found that not all strains habituate at the same rate. B6 decreased their locomotor activity successfully in both a five and a fifteen minute trial (present study, Fig. 1; Logue et al., 1997; Bolivar, 2009) whereas A mice decreased locomotion in a 15-minute trial (Bolivar, 2009) but not when assessed in a five minute trial (present study, Fig. 1; Logue et al., 1997). Similarly, avoidance behavior remained stable in the present study (Fig. 1), but this behavior increased between the first five minute epoch of a fifteen minute OF-trial and the total average of that trial in Gershenfeld and Paul (1997).

The lack of change in anxiety-related behavior and locomotion over time in A may thus have been (part) result of the relatively short trial length. Perhaps re-assessment this behavior over a longer period of time would produce a different picture of adaptive capacities in this strain.

Linkage studies aimed at the identification of chromosomal regions that modulate the change of behavior across time however suggest that the initial behavioral response is most closely linked to genotype. Radcliffe et al., (1998) for example identified different chromosomes on different time points when assessing locomotor activity in five-minute epochs of a 30-minute OF-trial, in a panel of long-sleep and short-sleep recombinant inbred strains of mice. They found that the coefficient of genetic determination decreased over the course of thirty minutes, leading them to suggest that an initial response to a novel environment is more dependent on genotype, whereas environmental influences become increasingly influential to behavioral output (Radcliffe

et al., 1998). Although our currently employed short trial length may thus have masked potential adaptive capacities in A mice, it does strengthen the argument that the observed differences between A and B6 are indeed related to genotype, and not environmental influences. Which in turn corroborates the purpose of this paper: chromosomal assignment of QTLs that modulate behavioral differences in the change of anxiety-related and activity behavior between these strains.

In the present consomic strain survey such QTL appeared located on mouse chromosomes 10 and 19 (Table 5). Chromosome 19 was associated with suggestive/significant evidence for one or more QTLs on anxiety-related behavior and avoidance behavior: High levels of overall anxiety-related behavior (Fig. 2), avoidance behavior (Fig. 3) and risk assessment (Fig. 4) in CSS-19 remained stable throughout the trial, resembling the temporal patterns displayed by A mice. Likewise, chromosome 10 was associated with significant evidence for one or more QTLs on risk assessment (Fig. 4). Both chromosomes have indeed repeatedly come forward as chromosomes harboring QTLs that modulate murine anxiety-related behavior, both in previous consomic surveys on (partly) the same mHB-data (Laarakker et al., 2008; Labots et al., 2016, see Table 5) as in other mapping populations with B6 and A as progenitor strains and other behavioral tests (F_2 intercross population and OF: Gershenfeld et al., 1997; F_2 intercross population and LD: Gershenfeld and Paul, 1997; CSS panel and LD: Singer et al., 2004; set of recombinant inbred strains or set of recombinant congenic strains and OF: Gill and Boyle, 2005; advanced intercross population or panel of interval-specific congenic strains and OF or LD: Zhang et al., 2005; CSS panel or F_2 intercross population and automated home cage environment: de Mooij-van Malsen et al., 2009 and de Mooij-van Malsen et al., 2013). Moreover, chromosome 19 has been linked to the temporal pattern of avoidance behavior in an F_2 intercross population based on B6 and CSS-19 (de Mooij-van Malsen et al., 2013), while significant/suggestive evidence for a role in the temporal aspect of anxiety-related behavior was found for both chromosomes in the OF and LD (Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Zhang et al., 2005).

At the same time, the present survey did not find evidence for other prominent chromosomes that were identified in previous analyses on this data, for example chromosome 15 and Y (Laarakker et al., 2008; Labots et al., 2016; see also Table 5). This however is in line with other studies that only found a partial overlap between the temporal component and overall expression of a

behavior. Gershenfeld and Paul (1997) found that sets of QTL affecting initial OF anxiety-related behavior (time spent in OF center in the first five minutes) only partially overlap with QTLs found for habituated OF behavior (measured as the difference between the first five minutes and the last five-minute epoch of a 15 minute trial). Similarly, Gershenfeld et al., (1997) found partial overlap in OF-locomotor activity between the first five minutes of a first trial, and the last five-minute epoch of a second trial that was recorded after a two week interval. In the survey by Labots et al., (2016) chromosome 19 came forward as especially interesting as this chromosome was the only chromosome that modulated anxiety-related behavior, but not locomotor activity. This was considered promising because locomotion, as described earlier in this section, may confound the expression of anxiety-related behavior (Ohl, 2003).

Like anxiety-related behavior, locomotor activity in mice has been found to be polygenic, with the majority of mouse chromosomes being identified to carry QTLs affecting locomotor activity (Radcliffe et al., 1998). This polygenic nature was also found for the temporal aspect of locomotion as significant/suggestive evidence was found in 11 of the 20 tested consomic lines when analyzing the AUC combined with latencies (Fig. 9, Table 5). Moreover, analyzing the trajectories of locomotion revealed suggestive/significant evidence for one or more QTLs on *all* chromosomes (Fig.7, Table 5). The present survey as such partly supports the above described conclusion by Labots et al., (2016) considering chromosome 19. Assessing the change of behavior by means of the AUC combined with latencies revealed suggestive evidence in chromosome 19 for one or more QTLs that modulate anxiety-related behavior (Fig. 8) but not locomotion (Fig. 9). In contrast, chromosome 10 harbors one or more QTLs for both anxiety-related behavior and locomotion when analyzing the AUC combined with latencies. This might indicate one or more QTLs with a pleotropic effect, but this should be further explored by for example and F_2 intercross between CSS-10 and B6.

Also, strain differences in overall anxiety-related behavior between CSS-19 and the host strain remained intact after controlling for locomotion by including this behavior as a variable in the temporal analyses. This was the case for both the trajectory analyses and the AUC combined with latencies. This strain difference disappeared however for CSS-10. The only exception to this pattern were the analyses of the trajectory of locomotion, i.e. the z-transformed variable 'total number of line crossings', in which evidence for a QTL for anxiety-related behavior (i.e. risk assessment) and locomotion was found for both chromosome

10 and 19 (Fig. 4, 7). In the case of mouse chromosome 19, this could suggest that this chromosome carries QTLs that affect the change of the number of line crossings over time, but not the total number of line crossings displayed in a single mHB-trial (Table 5).

An alternative explanation for this phenomenon may however simply be that transposing the single-trial data to a trajectory may allow for the identification of more subtle differences between the consomic lines and the host strain. In all consomic lines, low levels of this behavioral variable increased, while B6 decreased their initially high number of line crossings over the course of five minutes (Fig. 7). In the majority of consomic lines, initial strain differences in locomotor activity with B6 disappeared towards the end of the trial. The survey by Laarakker et al., (2008) described how a large number of chromosomes affected the average expression of the number of line crossings (n = 13/21) but found no evidence for (a) QTL(s) for this behavior in CSS-2, CSS-7, CSS-8, CSS-9, CSS-14, CSS-17, CSS-19 and CSS-Y. In the present survey, differences between B6 and these particular lines were only found in the first epoch, whereas in other lines the differences persisted up to the second, third, fourth epoch, or remained lower on all five epochs (Fig. 7). The differences between these lines and B6 as such appear less pronounced and it is possible that these differences were not picked up when collapsing the total scores in a single value.

Finally, like locomotion, exploration acts as a counterpart of expressed anxiety (Ohl, 2003): Behavior displayed in a novel environment/stimulus is often regarded as the net result of a conflict between the drive to explore versus the motivation to avoid potentially harmful stimuli (the approach/avoidance conflict, e.g. Ohl, 2003). Once a novel stimulus is assessed to be safe, anxiety-related behavior typically decreases and exploration increases. This interplay was indeed confirmed by previous analyses on this data, in which for example the variable '*total number of board entries*' (avoidance behavior) correlated strongly with '*total number of hole explorations*' (exploration, r = 0.940), and both these variables loaded highly on the same component (Factor 2) in a factor analysis (Tables 3 and 8 respectively in Laarakker et al. 2008). Given this interplay, it would be equally interesting to identify chromosomes that act on anxiety-related behavior but not on exploration. In the present study, controlling for exploration by including this variable as a covariate in the anxiety analyses in the consomic strain survey yielded inconclusive results. The evidence for a meaningful difference in anxiety between CSS-19 and B6 remained after controlling for exploration in the trajectory analyses (GLMM),

Table 4. Type of evidence for the GLMMs, based on Cohen's *d* and *P*-values.

Omnibus test		***Post hoc* comparison**	**Evidence for a meaningful QTL/ Meaningful difference between host and donor strain**
Strain effect (S)	Interaction Strain x Epoch	Epoch 1, 2, 3, 4, 5[a]	
$P > 0.05$	$P > 0.05$	$\|d\| < 1.0$	No
$P > 0.05$	$P > 0.05$	$1.0 \leq \|d\| < 1.5$ (#)[b]	No
$P > 0.05$	$P > 0.05$	$\|d\| \geq 1.5$ (##)	No
$P > 0.05$	$0.004 > P \leq 0.05$ (*)	$\|d\| < 1.0$	No
$P > 0.05$	$0.004 > P \leq 0.05$ (*)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P > 0.05$	$0.004 > P \leq 0.05$ (*)	$\|d\| \geq 1.5$ (##)	Suggestive
$P > 0.05$	$P \leq 0.004$ (**)	$\|d\| < 1.0$	No
$P > 0.05$	$P \leq 0.004$ (**)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P > 0.05$	$P \leq 0.004$ (**)	$\|d\| \geq 1.5$ (##)	Significant
$0.004 > P \leq 0.05$ (*)	$P > 0.05$	$\|d\| < 1.0$	No
$0.004 > P \leq 0.05$ (*)	$P > 0.05$	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$0.004 > P \leq 0.05$ (*)	$P > 0.05$	$\|d\| \geq 1.5$ (##)	Suggestive
$0.004 > P \leq 0.05$ (*)	$0.004 > P \leq 0.05$ (*)	$\|d\| < 1.0$	No
$0.004 > P \leq 0.05$ (*)	$0.004 > P \leq 0.05$ (*)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$0.004 > P \leq 0.05$ (*)	$0.004 > P \leq 0.05$ (*)	$\|d\| \geq 1.5$ (##)	Suggestive
$0.004 > P \leq 0.05$ (*)	$P \leq 0.004$ (**)	$\|d\| < 1.0$	No
$0.004 > P \leq 0.05$ (*)	$P \leq 0.004$ (**)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$0.004 > P \leq 0.05$ (*)	$P \leq 0.004$ (**)	$\|d\| \geq 1.5$ (##)	Significant
$P \leq 0.004$ (**)	$P > 0.05$	$\|d\| < 1.0$	No
$P \leq 0.004$ (**)	$P > 0.05$	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P \leq 0.004$ (**)	$P > 0.05$	$\|d\| \geq 1.5$ (##)	Significant
$P \leq 0.004$ (**)	$0.004 > P \leq 0.05$ (*)	$\|d\| < 1.0$	No
$P \leq 0.004$ (**)	$0.004 > P \leq 0.05$ (*)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P \leq 0.004$ (**)	$0.004 > P \leq 0.05$ (*)	$\|d\| \geq 1.5$ (##)	Significant
$P \leq 0.004$ (**)	$P \leq 0.004$ (**)	$\|d\| < 1.0$	No
$P \leq 0.004$ (**)	$P \leq 0.004$ (**)	$1.0 \leq \|d\| < 1.5$ (#)	Suggestive
$P \leq 0.004$ (**)	$P \leq 0.004$ (**)	$\|d\| \geq 1.5$ (##)	Significant

[a] The largest Cohen's *d* effect size of all five epochs was used to determine the evidence for a meaningful effect. [b] The symbols in parentheses (*, **, #, ##) are also used in Figures 1, 2, 3, 4, 5 and 6.

but not when analyzing the AUC combined with latencies (ANCOVA). The suggestive evidence for a meaningful difference in anxiety between CSS-10 and B6 when analyzing the AUC combined with latencies disappeared after controlling for exploration. These results could indicate a genetic basis for the interplay between exploration and anxiety, but further research is necessary to disentangle this relationship.

5. Conclusion

All in all, the present survey suggests that mouse chromosome 19 not only modulates the expression of overall anxiety-related behavior, but also the temporal component of this behavior in the mHB. Evidence for a similar modulating effect was found on chromosome 10, but this effect disappeared after controlling for a potentially confounding effect of locomotion and exploration.

Table 5. Suggestive and significant evidence for QTLs influencing anxiety-related behavior and/or activity behavior in the mHB of male inbred mice[a]

MOTIVATIONAL SYSTEM / *Behavioral dimension* / *Behavioral variables*	**Chromosomes**																				**Type of analysis[b]**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**X**	**Y**	
	QTLs influencing the difference in mHB behavior between male B6 and A mice																					
Locomotion AUC	**X**	–	**X**	na	–	**X**	–	–	–	**X**	x	**X**	x	–	**X**	**X**	–	**X**	–	x	–	A
Locomotion	**X**[c]	–	x	x	–	**X**	–	–	–	–	x	–	–	–	**X**	–	–	**X**	–	X	–	Z
Total number of line crossings	**X**	–	**X**	**X**	x	**X**	–	–	–	**X**	**X**	**X**	**X**	–	**X**	**X**	–	**X**	–	X	–	U
Total number of line crossings *trajectory*[e]	**X**	**X**	**X**	na	x	**X**	**X**	**X**	x	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	T
Latency until the first line crossing	x	–	–	–	–	**X**	–	–	x	x	–	x	–	–	x	–	–	**X**	–	–	–	U
Factor 1 – DI/ME/LO[d]	**X**	–	x	–	–	**X**	x	–	–	**X**	x	x	x	–	–	–	–	**X**	–	–	–	F
ANXIETY corrected for *Locomotion*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	U
ANXIETY corrected for *Locomotion trajectory*	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	T
ANXIETY corrected for *Locomotion AUC*	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	A
ANXIETY	x	–	–	–	–	–	–	–	–	x	–	–	–	–	x	–	–	–	x	X	–	U
ANXIETY *trajectory*	–	–	–	na	–	–	–	–	–	x	–	–	–	–	–	–	–	–	**X**	–	–	T
ANXIETY *AUC*	–	–	–	na	–	–	–	–	–	x	–	–	–	–	–	–	–	–	x	–	–	A
Avoidance	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	**X**	–	x	Z
Avoidance trajectory	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	T
Avoidance AUC	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	A
Total number of board entries	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**X**	–	–	U
Latency until the first board entry	–	–	–	–	x	–	–	–	–	–	–	–	–	–	–	–	–	–	**X**	–	x	U
Percentage of time on the board	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Average duration of a board entry	–	–	–	–	–	–	–	x	–	**X**	–	–	–	–	–	–	–	–	–	–	–	U

MOTIVATIONAL SYSTEM / *Behavioral dimension* / *Behavioral variables*	**Chromosomes**																				**Type of analysis[b]**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**X**	**Y**	
	QTLs influencing the difference in mHB behavior between male B6 and A mice																					
Factor 2 – AV/UN/OT	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	F
Risk assessment	x	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	x	–	–	Z
Risk assessment AUC	–	–	–	na	–	–	–	–	–	x	–	–	–	–	–	–	–	–	x	–	–	A
Total number of stretched attends	x	–	–	–	–	–	–	–	–	**X**	–	–	–	–	**X**	–	–	x	**X**	x	–	U
Total number of stretched attends *trajectory*[e]	–	–	–	na	–	–	–	–	–	**X**	–	–	–	–	–	–	–	–	x	–	–	T
Latency until the first stretched attend	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–	–	**X**	x	–	U
Factor 7 – RI/UN	**X**	–	–	–	–	x	–	–	–	**X**	–	x	–	–	**X**	–	–	x	x	**X**	–	F
Arousal	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–	–	–	–	–	Z
Arousal trajectory	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T
Arousal AUC	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	A
Total number of self-groomings	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Latency until the first self-grooming	–	–	–	–	–	X	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Percentage of time self-grooming	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Average duration of a self-grooming	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Latency until the first self-grooming + Percentage of time self-grooming	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	B
Latency until the first self-grooming + Average duration of a self-grooming	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	B
Total number of defecations	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–	–	–	–	–	U
Latency until the first bolus	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Total number of urinations	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–	–	–	–	–	U
Latency until the first urination	–	–	–	–	–	–	–	–	–	x	–	–	–	–	–	–	–	–	–	–	–	U
Factor 3 – AR	–	–	–	–	–	**X**	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	F
Factor 4 – OT	–	–	x	–	x	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	**X**	F
Factor 6 – AR	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	F
Exploration trajectory	**X**	**X**	x	na	–	**X**	**X**	x	**X**	**X**	–	**X**	**X**	–	**X**	**X**	**X**	x	**X**	**X**	–	T
Exploration AUC	–	–	–	na	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	A
Total number of rearings in the box	x	–	–	–	–	**X**	**X**	–	–	**X**	–	**X**	–	–	**X**	x	–	x	–	x	–	U
Latency until 1st rearing in the box	**X**	–	–	–	–	**X**	–	–	–	**X**	–	–	–	–	–	–	–	–	–	–	–	U
Total number of rearings on the board	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Latency until 1st rearing on the board	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Total number of hole explorations	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Latency until 1st hole exploration	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Total number ofholes visited	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Latency until 1st hole visit	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	U
Factor 5 – DI	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	F
Factor 9 – DI/ME	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	x	–	–	–	–	F

[a] Based on n = 27 host and n = 6 consomic mice.

[b] Finding evidence for a QTL on a particular chromosome was based on different methods: T = trajectory analysis, integrated behavioral z-scoring (generalized linear mixed models) [this study]; A = areas under the curve analysis combined with latencies, integrated behavioral z-scoring (ANCOVA plus bootstrapped *P* values) [this study]; Z = effect size measurement, statistical significance testing (ANCOVA plus bootstrapped *P* values) and integrated behavioral z-scoring (Labots *et al.,* 2016); U = univariate statistical analysis (Laarakker *et al.,* 2008) (unpaired Student's *t* test, unpaired Student's *t* test with Welch-Satterthwaite correction, Wilcoxon-Mann-Whitney test); B = bivariate statistical analysis (Laarakker *et al.,* 2008) (Hotelling's T^2 test); F = Factor analysis (Laarakker *et al.,* 2008).

[c] **X** = significant, x = suggestive, and – = no evidence for a QTL on a particular chromosome, na = missing value.

[d] The factors are labelled with the behavioral dimensions they mainly reflect. Abbreviations used: AR = arousal, AV = avoidance , DI = directed exploration, LO = locomotion, ME = memory, OT = other behavior, RI = risk assessment, UN = undirected exploration.

[e] *z*-transformed variable used for analysis of behavioral trajectories for locomotion and risk assessment. Because latency variables were not included in composition of the behavioral z-scores for locomotion and risk assessment, the z-scores for these behavioral dimensions amounted to the z-score of the variables 'total number of line crossings' (locomotion) and 'total number of stretched attends' (risk assessment).

References

Baud, A., Flint, J. 2017. Identifying genes for neurobehavioral traits in rodents: progress and pitfalls. Dis. Models Mech. 10, 373-383, https://doi.org/10.1242/dmm.027789.

Belknap, J. K. 2003. Chromosome substitution strains: some quantitative considerations for genome scans and fine mapping. Mamm. Genome, 14, 723-732, https://doi.org/10.1007/s00335-003-2264-1.

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

Bland, J. M., Altman, D. G., Statistics notes: bootstrap resampling methods. BMJ, 350, h2622, https://doi.org/10.1136/bmj.h2622.

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, http://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.

Bolivar, V.J., Caldarone, B. J., Reilly, A. A., Flaherty, L., 2000. Habituation of activity in an open field: A survey of inbred strains and F1 hybrids. Behav. Genet. 30, 285-293, https://doi.org/10.1023/A:1026545316455.

Bothe, G.W.M., Bolivar, V.J., Vedder, M.J, Geistfeld, J.G., 2005. Behavioral differences among fourteen inbred mouse strains commonly used as disease models. Comp. Med. 55 (4), 326-334, PMID: 16158908.

Bouwknecht, J. A., Paylor, R. 2008. Pitfalls in the interpretation of genetic and pharmacological effects on anxiety-like behavior in rodents. Behav. Pharmacol., 19(5-6), 385-402, https://doi.org/10.1097/FBP.0b013e32830c3658.

Brooks, M. E., Kristensen K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., Bolker, B. M., 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modelling. R. J. 9 (2), 378-400.

Clément, Y., Calatayud, F., Belzung, C. 2002. Genetic basis of anxiety-like behaviour: A critical review. Brain. Res. Bull., 57 (1), 57-71, https://doi.org/S0361-9230(0)00637-2.

De Mooij-van Malsen, A. J. G., van Lith, H. A., Oppelaar, H., Hendriks, J., de Wit, M., Kostrzewa, E., Breen, G., Collier, D. A., Olivier, B., Kas, M. J. 2009. Interspecies trait genetics reveals association of Adcy8 with mouse avoidance behavior and a human mood disorder. Biol. Psychiatry 66 (12), 1123-1130, https://doi.org/j.biopsych.2009.06.016.

De Mooij-van Malsen, J.G., van Lith, H.A., Laarakker, M.C., Brandys, M.K., Oppelaar, H., Collier, D.A., Olivier, B., Breen, G., Kas, M.J. 2013. Cross-species genetics converge to TLL2 for mouse avoidance behavior and human bipolar disorder. Genes Brain Behav. 12, 653-657, https://doi.org/10.1111/gbb.12055.

Festing, M.F.W. 2014. Extending the statistical analysis and graphical presentation of toxicity test results using standardized effect sizes. Toxicol. Pathol. 42 (8), 1238-1249, https://doi.org/10.1177/0192623313517771.

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

Gershenfeld, H. K., Neumann, P. E., Mathis, C., Crawley, J. N., Li, X., Paul, S. M. 1997. Mapping quantitative trait loci for open-field behavior in mice. Behav. Genet., 27 (3), 201-210, https://doi.org/10.1023/a:1025653812535.

Gershenfeld, H. K., Paul, S. M. 1997. Mapping quantitative trait loci for fear-like behaviors in mice. Genomics, 46 (1), 1-8, https://doi.org/10.1006/geno.1997.5002.

Gill, K. J., Boyle, A. E. 2005. Quantitative trait loci for novelty/stress-induced locomotor activation in recombinant inbred (RI) and recombinant congenic (RC) strains of mice. Behav. Brain Res. 161, 113-124, https://doi.org/10.1016/j.bbr.2005.01.013.

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, http://doi.org/10.1016/j.jneumeth.2011.01.019.

Hartig, F. 2020. DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.3.3.0, https://CRAN.R-project.org/package=DHARMa.

Holmes, A. 2001. Targeted gene mutation approaches to the study of anxiety-like behavior in mice. Neurosci. Biobehav. Rev., 25 (3), 261-273, https://doi.org/10.1016/s0149-7634(01)00012-4.

Kas, M. J., de Mooij, A. J. G., Olivier, B., Spruijt, B. M., van Ree, J. M. 2008. Differential genetic regulation of motor activity and anxiety-related behaviors in mice using an automated home cage task. Behav. Neurosci., 122 (4), 769-776, https://doi.org/10.1037/0735-7044.122.4.769.

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 blood plasma and brain magnesium-ion concentration with anxiety. Phys. Behav. 102, 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, http://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, 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, http://doi.org/10.1016/j.jneumeth.2017.09.003.

Lander, E. S., Botstein, D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121 (1), 186-199, PMCID: PMC1203601.

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

Lovell, D.P. 2013. Biological importance and statistical significance. J. Agric. Food Chem. 61 (35), 8430-8348, https://doi.org/10.1021/jf401124y.

Logue, S.F., Owen, E.H., Rasmussen, D.L., Wehner, J.M., 1997. Assessment of locomotor activity, acoustic and tactile startle, and prepulse inhibition of startle in inbred mouse strains and F_1 hybrids: Implications of genetic background for single gene and quantitative trait loci analyses. Neurosci. 80 (4), 1075-1086, https://doi.org/10.1016/s0306-4522(97)00164-4.

Mancuso, S., 2020. Model-based bootstrapped AN(C)OVA, https://sammancuso.com/2020/02/06/model-based-bootstrapped-ancova/

Mathis, C., Neumann, P.E., Gershenfeld, H., Paul, S.M., Crawley, J.N. 1995. Genetic analysis of anxiety-related behaviors and responses to benzodiazepine-related drugs in AXB and BXA recombinant inbred mouse strains. Behav Genet 25, 557–568, https://doi.org/10.1007/BF02327579.

Moy, S. S., Nadler, J. J., Young, N. B., Perez, A., Holloway, L. P., Barbaro, R. P., Barbaro, J. R., Wilson, L. M., Threadgill, D. W., Lauder, J. M., Magnuson, T. R., Crawley, J. N. 2007. Mouse behavioral tasks relevant to autism: phenotypes of 10 inbred strains. Behav. Brain Res., 6 (1), 4-20, https://doi.org/10.1016/j.bbr.2006.07.030.

Nadeau, J.H., Singer, J.B., Matin, A.M., Lander, E.S. 2000. Analyzing complex genetic traits with chromosome substitution strains. Nat Genet 24, 221–225, https://doi.org/10.1038/73427.

Nadeau, J. H., Forejt, J., Takada, T., Shiroishi, T. 2012. Chromosome substitution strains: gene discovery, functional analysis, and systems studies. Mamm. Genome 23, 693-705, https://doi.org/10.1007/s00335-012-9426-y.

Nakagawa, S., Cuthill, I.C. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol. Rev. 82, 591-605, https://doi.org/10.1111/j.1469-185X.2007.00027.x.

Nakagawa, S., Schielzeth, H. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Meth. Ecol. Evol. 4 (2), 133-142, https://doi.org/10.1111/j.2041-210x.2012.00261.x

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

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. 2005. Animal models of anxiety. Handb. Exp. Pharmacol., (169), 35-69, https://doi.org/10.1007/3-540-28082-0_2.

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.

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lidster, K., MacCallum, C. J., Macleod, M., Petersen, O., Rawle, F., Reynolds, P., Rooney, K., Sena, E. S., Silberberg, S. D., Steckler, T., Würbel, H., 2020a. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. BMC Vet Res. 16, 242, https://doi.org/10.1186/s12917-020-02451-y.

Percie du Sert N., Ahluwalia A., Alam S., Avey M. T., Baker M., Browne W. J., Clark A., Cuthill I. C., Dirnagl U., Emerson M., Garner P., Holgate S. T., Howells D. W., Hurst V., Karp N. A., Lazic S. E., Lidster K., MacCallum C. J., Macleod M., Pearl E. J., Petersen O. H., Rawle F., Reynolds P., Rooney K., Sena E. S., Silberberg S. D., Steckler T., Würbel H. 2020. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLos Biol. 18 (7), e3000411, https://doi.org/10.1371/journal.pbio.3000411.

R Core Team, 2020. R: A language environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Radcliffe, R.A., Jones, B.C., Erwin. G. 1998. Mapping or provisional quantitative trait loci influencing temporal variation in locomotor activity in the LS x SS recombinant inbred strains. Behav. Genet. 28 (1), 39-47, https://doi.org/10.1023/a:1021456731470.

Rights, J.D., Sterba, S.K. 2019. Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. Psych. Meth. 24 (3), 309-338, https://doi.org/10.1037/met0000184.

Rogers, D. C., Jones, D. N. C., Nelson, P. R., Jones, C. M., Quilter, Ch. A., Robinson, T. L., Hagan, J. J., 1999. Use of SHIRPA and discriminant analysis to characterize marked differences in the behavioural phenotype of six inbred mouse strains. Behav. Brain Res. 105 (2), 207-217, https://doi.org/10.1016/s0166-4328(99)00072-8.

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, http://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, http://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, http://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, http://doi.org/10.1016/j.bbr.2013.02.023.

Singer, J. B., Hill, A. E., Burrage, L. C., Olszens, K. R., Song, J., Justice, M., O'Brien, W. E., Conti, D. V., Witte, J. S., Lander, E. S., Nadeau, J. H. 2004. Genetic dissection of complex traits with chromosome substitution strains of mice. Science, 304 (5669), 445-448, https://doi.org/10.1126/science.1093139.

Singer, J. B., Hill, A. E., Nadeau, J. H., Lander, E. S. 2005. Mapping quantitative trait loci for anxiety in chromosome substitution strains of mice. Genetics, 169 (2), 855-862, https://doi.org/10.1534/genetics.104.031492.

Tam, W. Y., Cheung, K-K., 2020. Phenotypic characteristics of commonly used mouse inbred strains. J. Mol. Med. Jul 25. https://doi.org/10.1007/s00109-020--1953-4.

Trullas, R., Skolnick, P. 1993. Differences in fear motivated behaviors among inbred mouse strains. Psychopharmacology, 111 (3), 323 – 331, https://doi.org/10.1007/BF02244948.

Turri, M.G., DeFries, J.C., Henderson, N.D., Flint, J. 2004. Multivariate analysis of quantitative trait loci influencing variation in zniety-related behavior in laboratory mice. Mamm. Genome, 15, 69-76, https://doi.org/10.1007/s00335-003-3032-y.

Van Gaalen, M. M., Steckler, T. 2000. Behavioral analysis of four mouse strains in an inxiety test battery. Behav. Brain Res., 115 (1), 95-106, https://doi.org/10.1016/S0166-4328(00)00240-0.

Van der Goot, M. H., Boleij, H., van den Broek, J., Salomons, A. R., Arndt, S. S., van Lith, H. A., 2020. An individual based, multidimensional approach to identify emotional reactivity profiles in inbred mice. J. Neurosci. Meth. https://doi.org/10.1016/j.neumeth.2020.108810.

Wahlsten, D., 2011. Sample size. In: Mouse behavioral testing: how to use mice in behavioral neuroscience, 1st edn., Academic Press, Elsevier Inc., London, U.K., pp 75-105.

Zhang, S., Lou, Y., Amstein, T. M., Anyango, M., Mohibullah, N., Osoti, A., Stancliffe, D., King, R., Iraqi, F., Gershenfeld, H. K. 2005. Fine mapping of a major locus on chromosome 10 for exploratory and fear-like behavior in mice. Mamm. Genome 16 (5), 306-318, https://doi.org/10.1007/s00335-00402427-8.

General discussion

Aim

The overarching aim of this thesis was twofold: *i)* to gain insight in how inter-individual variability in adaptive capacities in relation to anxiety is expressed within and between multiple mouse inbred strains; and *ii)* to improve our understanding of how such inter-individual variability can be incorporated in statistical analysis and experimental design, and how this affects the quality of experimental results. The adaptive quality of anxiety responses was measured as the behavioral and physiological (corticosterone) response in multiple mouse inbred strains, during repeated exposure to a mild aversive stimulus (the modified Hole Board, abbreviated to mHB). We used an unsupervised multivariate clustering approach that was designed specifically for longitudinal response trajectories to analyze behavioral and physiological response trajectories of mice on an individual level. With this clustering procedure we assessed whether we could identify subgroups of mice that followed the same response over multiple behavioral dimensions: response types.

Application of this approach when re-analyzing existing mHB data on behavioral habituation of anxiety-related behavior in BALB/cJ and several sub-strains of 129 mice yielded two subtypes of response of differential adaptive value (**Chapter 2**). These retrospect analyses as such provided a first indication that adaptive capacities may indeed differ within inbred strains: in 129 mice, the majority was characterized by impaired habituation of anxiety-related behavior, while a subgroup of 129 individuals in fact displayed successful habituation when repeatedly exposed to the mHB. In addition, this study indicated that defining experimental animals on an individual level may not only yield new information regarding to subtypes of response, but also may yield new information regarding the adaptive quality of these differential responses. Inter-individual variability in adaptive capacities in 129S2/SvPasCrl (129S2) mice was empirically confirmed in two follow-up experiments. In these experiments however, differential subtypes of response were now also found in BALB/cAnNCrl (C), and another commonly used mouse inbred strain, C57BL/6NCrl (B6N) (**Chapters 3 & 4**). The two response types that resulted from the same clustering procedure in these experiments were similar to the differential profiles found in **Chapter 2**, and were displayed by individuals of all three strains. These two studies furthermore showed that individuality in adaptive capacities was reflected in behavior only, and not on a physiological level: In both studies, the identified subtypes differed significantly on anxiety-related and activity behavior, but not on circulating corticosterone (pCORT) levels (**Chapters 3 & 4**).

With respect to the quality of experimental results, we empirically demonstrated how inter-individual variability alters the interpretation of a pharmacological experiment when evaluating the effects of an anxiolytic compound on behavior (**Chapter 4**). Systematic incorporation of the two individual response types in the design of a pharmacological experiment produced different results from an experimental pool in which this information was not accounted for: Treatment effects were augmented while confounding experimenter effects were masked by inter-individual variability (**Chapter 4**). In addition to zooming in on individual responses, this thesis provides a comprehensive study of behavioral and physiological phenotypes of the change of anxiety-related and activity behavior between three mouse inbred strains that differ in innate emotionality (**Chapter 3** & **Box II**). Finally, a third objective of this thesis was to investigate the genetic underpinnings that modulate inter-individual variability in habituation of anxiety responses. We applied a genetic strategy to identify quantitative trait loci that modulate habitation of anxiety responses and showed that some of the chromosomes associated with overall anxiety-related behavior in the mHB are also associated with the change of anxiety-related behavior over time (**Chapter 5**).

The following sections first discuss the identified response types and their implications in more detail. This is followed by a discussion of the findings regarding the incorporation of inter-individual variability in statistical analysis and design and their implications. Next, the limitations of this thesis in relation to the findings are addressed, as well as a number of methodological considerations. Lastly, this chapter concludes with suggestions for further research and a summarizing statement.

Inter-individual variability in habituation of anxiety: two response types

An important contribution of this thesis is that it provides an in-depth characterization of inter-individual variability in habituation of anxiety in response to an unprotected environment within, and between, three mouse inbred strains. The multivariate clustering procedure that assessed this inter-individuality identified two clusters throughout this thesis. These response types were predominantly characterized by contrasting patterns of avoidance behavior, and differences in exploration and locomotor activity. In addition, their behavioral profiles were highly similar between Chapters 2, 3 and 4. The sections below provide a summary characterization of these response types.

Summary characterization of response types

As outlined in Chapter 1, we assessed inter-individuality in habituation of anxiety-related behavior we expanded on a series of previous studies that were conducted to assess whether enhanced anxiety during repeated exposure to a stressor could mirror a non-adaptive anxiety response (Ohl et al., 2008) that may ultimately be employed as a means to operationalize pathological anxiety in mouse models (Boleij et al., 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). These studies compared the adaptive quality of anxiety responses by repeatedly exposing two mouse inbred strains (C and sub-strains of 129 mice) to the mHB. The response of C to such repeated exposure was characterized by initial high levels of anxiety that decreased over trials, while activity behavior increased. In contrast, in 129 mice anxiety-related behavior consistently increased, whereas activity behavior remained low or decreased over trials. In the present thesis, these differential behavioral profiles were also observed in C and 129S2 mice (Chapter 3, Box II). Salomons et al. (2012) furthermore demonstrated that these differential responses were sensitive to pharmacological manipulation in C and 129P3/J (129P3) mice. In addition, neuronal activity differed between C and 129P3 in brain regions associated with cognitive and emotional processing, where impaired neural activity was demonstrated in 129P3 but not in C (Salomons et al., 2010a; 2010c; 2013). On the basis of these combined results the authors concluded that the response of C might be classified as an adaptive anxiety response, whereas the response of 129 mice could be classified as non-adaptive anxiety.

One of the primary aims of this thesis was to evaluate to what extent the adaptive quality of these anxiety responses may differ between individuals of C and 129S2 mice, and to what extent such inter-individual variability might be present in a third strain, B6N. This yielded two differential multidimensional response types, which were displayed by individuals of all three strains. Interestingly, these two behavioral profiles bared a strong resemblance to the differential behavioral profiles displayed by C and 129 mice that were originally observed in the studies by Boleij and Salomons and colleagues (Chapter 2, 3 and 4). In contrast to the studies by Salomons et al. however, the present assessment was restricted to mapping inter-individual variability on a behavioral and physiological level, and did not include (successful) pharmacological validation or assessment of the brain-behavior relationship that may underlie this individuality. Whether these subtypes of response actually are indicative of (a lack of) adaptive capacities therefore requires further evaluation, which is discussed further on in this chapter.

In any event, in each of these chapters, the two response types were referred to as A and B, where A was always the larger cluster. Which of the two response types formed the largest cluster however differed between chapters, and as a result referring to response type A and B would not be suitable here. Given the resemblance in behavioral profiles, and to avoid confusion, we will therefore refer to the two individual response types as 'adaptive anxiety' versus 'non-adaptive anxiety' in the remainder of this chapter, with the side note that further research is necessary to confirm any differential adaptive quality between these subtypes.

Table 1 presents an overview of the directionality of all behavioral dimensions in each cluster, for each chapter. Mice with an adaptive anxiety profile displayed a decrease in avoidance behavior, while exploration increased in all three chapters. In this response type, locomotion either increased (Chapters 2 and 4) or remained stable at a high level (in comparison to the other response type, Chapter 3). Risk assessment decreased in Chapters 2 and 3 while arousal remained stable across trials in Chapter 2 and increased in Chapter 3. Risk assessment and arousal were not included in the cluster analysis in Chapter 4 (see below). Conversely, mice with a non-adaptive anxiety profile increased avoidance behavior in all three chapters, while low levels of locomotion remained stable (Chapters 2 and 3) or decreased across trials (Chapter 4). The patterns of exploration in this cluster were more diverse, with a decrease in Chapter 2, stable low levels of exploration in Chapter 4, and an increase in Chapter 3 (Table 1). Risk assessment decreased, while arousal increased or remained stable in chapters 2 and 3 (Table 1).

Table 1. Schematic overview of directionality of behavioral dimensions during repeated exposure to the mHB, for each response type, and for each chapter.

	Response type					
	Adaptive anxiety			**Non-adaptive anxiety**		
	Chapter 2	**Chapter 3**	**Chapter 4**	**Chapter 2**	**Chapter 3**	**Chapter 4**
Avoidance behavior	↓	↓	↓	↑	↑	↑
Risk assessment	↓	↓	-	↓	↓	-
Arousal	→	↑	-	→	↑	-
Exploration	↑	↑	↑	↓	↑	→
Locomotion	↑	→	↑	→	→	↓

Arrows indicate the directionality of the change of behavior during repeated exposure to the mHB.
↓ decrease;↑ increase; → no change in behavior during the first and the last mHB trial.

In all chapters, the two clusters were relatively close in size. In Chapters 2 and 4, the non-adaptive anxiety type formed the largest cluster (Chpt. 2, 58.4%; Chpt. 4, 57.5%) while in Chapter 3 mice of the adaptive anxiety type dominated the experimental pool (53.8%). Also, within-strain variability in response types differed between strains. In all chapters the majority of C mice grouped together in the adaptive anxiety cluster (respectively 100%, 82.5% and 88.1%). In contrast, the majority of 129S2 mice gravitated towards the non-adaptive anxiety type in the three chapters (respectively 85.8%, 76.3% and 90.0%). B6N mice were only assessed in Chapters 3 and 4 and similar to 129S2 gravitated towards the non-adaptive anxiety type, although the within strain variability was higher than in C and 129 mice (respectively 53.8% and 70.0%). These strain differences in within-strain variability are in line with previous reports of a relatively stable phenotype in C (low within-strain variability) compared to 129S1, but not in line with the fact that C57BL/6 (B6) mice have been equally characterized as having a stable phenotype (Loos et al. 2015).

Multidimensionality of individual response types

Anxiety in humans is a heterogeneous phenomenon with a complex inheritance, which involves the interaction of multiple gene variants with environmental and epigenetic factors (Hettema et al., 2001; Merikangas and Pine, 2002; Cryan and Holmes, 2005; Young et al., 2008). A similar polygenetic, epistatic and epigenetic nature has been demonstrated for anxiety-related behavior in mice (Turri et al., 2001; Lathe, 2004; Cryan and Holmes, 2005). This complex nature is also reflected on a behavioral level, where unconditioned anxiety is expressed by a combination of anxiety-related and activity behaviors (Belzung and Griebel, 2001; Ohl, 2003). This thesis therefore used a multidimensional test to assess inter-individual variability in habituation of anxiety responses, and a multivariate clustering procedure to characterize the individual mice.

A methodological advantage of multidimensional assessment over a univariate approach is that multiple motivational systems may be studied in parallel (Ohl, 2003). This advantage was indeed evident in the behavioral profiles of the response types in this thesis, which directly reflected the complex interplay between avoidance behavior, exploration and locomotion that is often observed in tests that measure unconditioned responses to a novel environment (Ohl, 2003; Ohl, 2005; O'Leary et al., 2013).

Avoidance behavior is often regarded as one of the most prominent indicators of anxiety (Barnett, 1975; Belzung and Le Pape, 1994), especially in behavioral tests that capitalize strongly on the so-called approach-avoidance conflict that arises when rodents are exposed to a novel/unprotected environment (such as the mHB; Ohl, 2003). This was confirmed by the fact that in all three chapters, avoidance behavior exerted the most 'weight' in partitioning of the response types (Chapter 2, 3, 4). An overlapping and prominent feature of the identified response types across chapters was that they were characterized by contrasting (increase *vs*. decrease) patterns of avoidance behavior (Table 1). In the original experiments that were analyzed in retrospect in Chapter 2, avoidance behavior was the most prominent indicator of impaired habituation in comparison to other anxiety-related behaviors such as arousal and risk assessment behavior (Boleij et al., 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). In C, B6N and 129S2 mice, avoidance behavior patterns differed significantly between strains (Chapter 3, Box II), while between strain differences in the other two anxiety-related dimensions, risk assessment and arousal, were less pronounced. When taking these factors together, the relative importance of avoidance behavior in determination of the differential response types seems hardly surprising.

As outlined in various sections of this thesis, the aforementioned approach-avoidance conflict was also reflected in the differential patterns of exploration that were found between response types. Exploration is often considered an indirect measure of anxiety (O'Leary et al., 2013) as exploratory behaviors may be gradually inhibited by anxiety during exposure to a novel environment (Ohl 2003). In Chapters 2, 3, and 4 this interplay was indeed found in the response types that reflected successful habituation, where a decrease in avoidance behavior was coupled to an increase in exploration (Table 1). Conversely, mice in the non-adaptive anxiety cluster combined an increase in avoidance behavior with a decrease in exploration in Chapters 2 and 4 (Table 1). In Chapter 3, this interplay was less pronounced (Table 1), as an increase in avoidance behavior was coupled with an increase in exploration – although the increase of exploration in this cluster was less pronounced than the increase in exploration in the adaptive anxiety cluster (Chapter 3).

Similar to exploration, locomotion is not only associated with activity levels but may also confound a correct interpretation of anxiety-related behavior. This confounding effect mostly pertains to avoidance behavior, as the lack of exploration of a novel environment may just as well be the result of low activity,

instead of an expression of anxiety (Ohl, 2005). It is of note however that other anxiety-related variables such as risk assessment and defecations have also been associated with differential locomotion levels (Milner and Crabbe, 2008; O'Leary et al., 2013). In addition to differential patterns of avoidance behavior, the clusters were also characterized by differential locomotion levels: individuals that decreased avoidance behavior either increased locomotion over time (Chapters 2 and 4), or displayed overall higher levels of locomotion compared to their counterparts in the cluster that increased avoidance behavior (Chapter 3). In contrast, overall locomotion was stable at lower levels (Chapters 2, 3) or decreased across trials (Chapter 4) in the clusters that were characterized by an increase in avoidance behavior. These response types as such provide a comprehensive insight in how inter-individuality in the interplay between anxiety-related and activity behavior may be expressed within C, 129S2 and B6N mice.

The multivariate analysis equally demonstrated, how some dimensions appeared more influential in determination of the clusters than others (Chapters 2, 3 and 4). Most notably, risk assessment, arousal and pCORT did not markedly differ between the two response types across chapters. These findings were somewhat surprising as individuality has been demonstrated for both behavioral dimensions (Hager et al., 2014; Ducottet and Belzung, 2004) as well as pCORT (Sgoifo et al., 1996; Cockrem, 2013; Jakovcevski et al., 2008). Like avoidance behavior, risk assessment has been prominently associated with anxiety (Blanchard et al. 2011). In fact this behavior has been suggested to be a more sensitive indicator of anxiety than avoidance behavior (i.e. Grewal et al. 1994; Rodgers and Cole 1994; Ohl et al. 2008). The expression of this behavior however is also highly strain specific (O'Leary et al. 2013). In this thesis, risk assessment was significantly lower in B6N compared to C and 129S2 (Chapter 3, Box II), which corroborates previous reports of low risk assessment behavior in B6 mice (Lepicard et al. 2000; Augustsson and Meyerson 2004; O'Leary et al. 2013). Also, risk assessment did not significantly differ between C and 129S2 mice in Chapter 3, and only differed on the first trial in Box II, with significantly higher risk assessment in C compared to 129S2. These findings are in line with previous reports of risk assessment behavior in C and 129 mice in the mHB, which demonstrated either no difference between strains (Salomons et al., 2010a; Salomons et al., 2013) or significantly higher risk assessment in C than 129 on the first trial (Salomons et al., 2010b). The overall variability of risk assessment behavior may therefore have been low in comparison to the variability in other behavioral dimensions, and as a result this variability may

has been conflated when pooling all behavioral dimensions in a single analysis. Cluster differences in risk assessment were indeed minimal in Chapter 2, while no significant cluster differences were observed in Chapter 3.

An alternative interpretation may however be that the lack of discriminative weight of the behavioral dimension risk assessment was inherent to the fact that we analyzed the temporal nature of this behavior. Throughout this thesis, risk assessment decreased across trials regardless of strain and/or cluster, which corroborates previous assessment of this behavior in the mHB (Boleij et al. 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). As a result, any variability in risk assessment behavior gravitated towards the first mHB trial(s), with little to none variability between individuals in the remaining trials. This pattern was also reflected in the fact that all models testing for between strain differences in risk assessment trajectories needed a correction for zero inflation (Chapter 3, Box II, Chapter 5). Rather than assessing inter-individual differences in the temporal pattern of this behavior, a more informational strategy would perhaps be to assess this variability on the first trial only.

Next, like risk assessment, behaviors comprising the behavioral dimension arousal, defecation and grooming-related variables, are known to be highly strain specific (O'Leary et al. 2013). Grooming duration for example, has been found to be low in C compared to B6 and 129 mice, while defecations were high in C compared to B6 and 129 mice (O'Leary et al. 2013). Mean strain values of these measures correlated highly across multiple tests of unconditioned anxiety, suggesting stable between strain differences (OF, EPM, LD; O'Leary et al. 2013). In Chapter 3, the composite behavioral z-score for arousal increased across trials in all three strains, but strain differences were small. In Box II, strain differences were more pronounced, with arousal increasing in B6N and 129S2, but not in C, and overall levels of arousal ranking highest in B6N, followed by C, and 129S2. These inconsistent results may be explained by the fact that both defecation and grooming have been linked to different processes. While both behaviors have been associated with anxiety in mice (Henderson et al. 2004; Estanislau 2012), demonstrations of defecation as an indicator of anxiety have been inconsistent (Turri et al. 2000; Kalueff and Tuohimaa, 2005a). Also, grooming can increase in both high stress and low stress environments, and the sequence of grooming patterns differs between the two (Kalueff and Tuohimaa, 2005a). Inbred strains may even differ in their sequence of grooming behaviors, with 129S1 (not tested here) displaying disrupted grooming patterns compared

to C mice (Kalueff and Tuohimaa, 2005b). The fact that the mHB-ethogram in our studies did not distinguish between stress-related and non-stress related grooming behavior makes the generally observed increase in arousal across strains difficult to interpret. Perhaps a distinction between these types of grooming would have resulted in more pronounced strain differences and in turn would have allowed for a better interpretation of inter-individual variability in both forms of arousal. In its current form however, the variability between individuals in arousal was minor compared to the variability in other behavioral dimensions. Arousal trajectories were stable over time in both clusters in Chapter 2, with no meaningful difference between clusters. Not surprisingly, omitting arousal from the cluster analysis did not have any effect on the distribution of mice across the two clusters in this chapter. In Chapter 3, arousal trajectories increased in both clusters with a more pronounced increase of arousal in the non-adaptive anxiety cluster. Omitting this dimension from the cluster analysis however resulted in 99% of the mice retaining their cluster, again suggesting little to no effect of this dimension on the cluster partitioning.

Taking together the above considerations, we suspect that, rather than the absence of any inter-individual variability, the overall variability of risk assessment and arousal was likely low in comparison to the variability in avoidance behavior, exploration and locomotion. As a result, this variability may have been conflated when pooling all behavioral dimensions in a single analysis. It has indeed been shown that *k*-means clustering approaches (such as *kml3d* applied in our analyses) may exert a bias towards identified equal sized clusters, which may result in behavioral dimensions exerting more subtle differences being hidden among larger groups (Kiselev et al. 2019). In single-cell RNA sequencing for example, in which clustering approaches are frequently applied to identify putative cell types, rare cell types may remain hidden in larger groups (Kiselev et al., 2019). In Chapters 2 and 3 all dimensions were included in the clustering procedure because it was either the first time we applied this multivariate clustering approach on mHB-data (Chapter 2) or because it was the first time we applied it in a controlled experiment, including pCORT (Chapter 3). In Chapter 4, arousal, risk assessment and pCORT were initially included in the cluster analysis but omitted at a later stage because the contribution of these 3 parameters to the partitioning of the clusters was relatively low in comparison to avoidance behavior, exploration and locomotion. Leaving these two dimensions out did not significantly change the cluster composition compared to inclusion of these dimensions ($\chi^2_1 = 0.82$, $P = 0.389$), while 95% of the mice ($n = 171/179$) retained their cluster. The goal here was to fine-tune the behavioral profiles that we would ultimately use to match the mice in the subsequent pharmacological part of the study. This procedure is also referred to as feature selection, a strategy that is often employed in clustering techniques with the goal to yield more compact clusters that are easier to interpret (Frades and Matthiesen 2010).

All in all, these findings illustrate how a potential risk of a multivariate approach is that more subtle effects may become conflated when pooling multiple variables into a single analysis and that researchers should be mindful of which variables are included when pooling multiple variables into a single analysis. When the goal is to investigate more subtle effects, a uniform clustering approach may be more appropriate. Yet, our results also demonstrate that particularly avoidance behavior trajectories, as well as patterns of exploration and locomotion may differ between individual C, 129S2 and B6N mice, and that multivariate assessment of these behavioral dimensions produces subtypes of response that are similar between studies.

Physiological indicators of inter-individual variability in anxiety

In the two experimental studies, the behavioral assessment of inter-individuality in anxiety phenotypes was supplemented with the assessment of pCORT (Chapter 3, Box II). In rodents, pCORT levels are often used as an indicator of acute stress, and have been associated with anxiety-related behavior (Korte et al., 2001; Ardayfio and Kim, 2006, Rodgers et al., 1999). In addition, multiple studies indicated that pCORT levels may vary greatly between individuals (Sgoifo et al., 1996; Rougé-Pont et al., 1998; Cockrem, 2013; Ebner and Singewald, 2017; Weger and Sandi, 2018). pCORT levels also varied substantially between 129 sub-strains assessed in the mHB (Boleij et al., 2012). It was therefore surprising to find that the identified response types were not reflected in differential profiles of circulating corticosterone levels, with no significant differences in pCORT between clusters (Chapter 3).

It's possible that the relative lack of discriminative weight of pCORT in the partitioning of the clusters was related to a relative lack in inter-individual variability, similar to risk assessment and arousal. It is of note however that in Chapter 3, the relative weight of pCORT on the clustering procedure was higher than exploration and locomotion, with 92% of the individuals retaining their cluster against 96% for exploration and locomotion, which would contradict this explanation. Also, between strain differences in pCORT were quite pronounced in both experiments (Chapter 3, Box II). pCORT levels were lowest in B6N compared to 129S2 and C mice on all time points, in both studies, which

is in accordance with previous assessment of this strain in the mHB (Ohl et al., 2001b). pCORT levels were highest in 129S2 compared to C and B6N directly after the behavioral test, which corroborates previous assessment of 129 mice in the mHB (female 129P3, Salomons et al., 2010a). Because between-strain differences were most pronounced directly after the mHB test we additionally explored whether the two clusters in Chapter 3 perhaps differed when only assessing this difference on sampling moment two, but this was not the case ($P = 0.1427$).

A definitive explanation for these results is difficult to provide, as pCORT levels are not only associated with anxiety, but may also differ under the influence of other factors, such as circadian rhythm (Kakihana and Moore, 1976), intestinal microbiome (Neuman et al., 2015), and nutritional status (Jensen et al., 2013). More studies are necessary to further explore whether the identified response types are also reflected on a physiological level. This may also include assessment of other physiological indicators that have been associated with anxiety and the expression of which differed between individuals, such as that of adrenocorticotrophic stress hormones (ACTH, Herman et al., 2016; de Boer et al., 2017), cardiovascular parameters (increased blood pressure and heart rate, Golbidi et al., 2015), and immune system responses (Menard et al., 2016). Such further assessment could contribute to a more definitive establishment of the adaptive quality of the identified response types.

Adaptive quality of individual response types

According to Armario and Nadal (2013) the in-depth characterization of inter-individual variability in habituation of anxiety responses provides a first step towards an improved understanding of any differential susceptibility towards the development of non-adaptive anxiety in the inbred strains under study. As noted above however, it cannot be determined at this stage whether these response types are indeed indicative of anxiety phenotypes with a differential adaptive quality, or whether they merely distinguish between individuals that are responsive to the aversive nature of the mHB (responders) versus individuals that are less responsive to the test (non-responders). Follow up steps, including pharmacological validation and further assessment of their underlying neurobiological mechanisms are necessary to determine to what extent these profiles indeed represent phenotypes of differential adaptive quality. In order to appropriately employ these steps however, a couple of factors should be taken into account.

The first is the aforementioned potentially confounding effect of locomotion. As described above, the clusters were characterized by contrasting patterns of avoidance behavior, but also by differential levels of locomotor activity (Chapters 2, 3, and 4). Mice that increased avoidance behavior typically displayed lower levels of locomotion, whereas mice that decreased avoidance behavior appeared more active. This interplay could indicate that the increase in avoidance observed in one portion of the mice was the result of low activity levels, instead of increased anxiety and vice versa (Boleij et al., 2012; O'Leary et al., 2013; Labots et al., 2016a). It is of note that multivariate analyses of mouse behavior have repeatedly demonstrated that anxiety displayed in response to novelty is not secondary to locomotion (mHB: Ohl et al., 2003; Laarakker et al., 2008; see Ramos and Mormède, 1998 for a literature review regarding other tests of unconditioned anxiety). At the same time, Labots et al. (2016a) demonstrated that differences in avoidance behavior between two sub-strains of B6 mice disappeared after controlling for locomotion in the analysis of avoidance behavior, suggesting that locomotion may exert confounding influences in the mHB. In Chapter 3, including the behavioral dimension locomotion as a covariate in the analysis did not change the observed cluster differences for avoidance behavior. Additional analyses for Chapters 2 and 4 showed the same effect (in both studies, trial effect: $P < 0.0001$; interaction cluster x trial: $P < 0.0001$), but the covariate locomotion did significantly contribute to the model in both studies ($P < 0.0001$). Further assessment of the adaptive quality of the identified response types would therefore benefit from a more definitive dissociation between activity and anxiety-related behavior.

A second factor that should be taken into consideration is to further evaluate the adaptive quality of the identified subtypes within each strain, rather than across strains. The inbred strains studied in this thesis are known for their differential anxiety phenotypes (Tam and Cheung, 2020), which were indeed confirmed in the two experiments conducted in this thesis (Chapter 3, Box II). These differential geno- and phenotypes bring with them their own intricate properties and questions which are specific for each strain. As described above, C mice in general are known for their rigid phenotype, showing less within-strain variability compared to for example 129 mice (Loos et al., 2015). The relatively small number of C that deviated from the majority of C mice (i.e. that displayed the opposite response type) may equally represent individuals that were less responsive to the test.

Similarly, it has been asked whether the impaired habituation pattern (i.e. an increase in avoidance behavior) in 129 mice reflects increased anxiety, or is the mere result of the reduced activity levels that are characteristic for this family of inbred strains (Cook et al., 2002; de Visser et al., 2006; Pratte and Jamon, 2009). The fact that pCORT levels were higher in 129S2 mice than in B6N and C directly after the behavioral test could indicate that exposure to the mHB was indeed perceived as particularly stressful by this strain. However, differential behavioral profiles in this study were not associated with differential pCORT levels (Chapter 3). Moreover, in Chapters 2, 3 and 4, an increase in avoidance behavior was indeed coupled with lower levels of locomotor activity – thus not precluding the possibility that a lack of exploration of the mHB was the result of low activity.

Finally, in B6N mice, two contrasting avoidance-behavior phenotypes emerged when analyzing the data on an individual level, while on average, low levels of avoidance behavior remained stable across trials in this strain. The low anxiety, high activity profile that is characteristic of B6 therefore warrants the question whether these two response types in B6N reflect differential anxiety profiles, or are the result of differential activity between B6N individuals. As discussed in Chapter 3, inter-individual variability in baseline anxiety-levels has been predominantly demonstrated in B6 mice (Ducottet and Belzung 2004; Jakovcevski et al., 2008; Lewejohann et al., 2011; Keshavarz et al., 2020). Previous research also indicated however that consideration of exploratory strategy may alter the interpretation of avoidance behavior in the mHB for this strain (Ohl et al., 2001b). B6 mice are known to demonstrate pronounced thigmotaxis when first exposed to a novel environment, meaning that they explore a novel test environment by staying close to the outer walls of an assay (Ohl et al., 2001b). This behavior was indeed also observed during data collection in both Chapters 3 and 4, and reflected in the relatively high locomotor activity (i.e. the number of line crossings and the latency to the first line crossing) for this strain in both chapters. The experiments in Chapters 3 and 4 did not evaluate whether the contrasting patterns of avoidance behavior were perhaps associated to subtle differences in exploration strategy. Such subtle differences could for example constitute inter-individual variability in the distance B6N mice kept towards the outer wall of the mHB while exploring the environment. Perhaps some individuals remained closer to the wall while others performed this strategy with a more central circle (and thus briefly crossing the unprotected area). To assess this possibility in hindsight we conducted a quick retrospect scan of the videos taken from the first trials of the majority (n = 34, 85%) of B6N mice tested in Chapter 3 to assess whether variability

in exploration strategy could have indeed formed a distinguishing factor for this strain between the two response types of this study. By superimposing a line on these videos separating the outer half of the protected area (i.e. the box) from the inner half of the protected area, we scored the time spent in each half (inner/outer) to assess whether individuals demonstrating low levels of avoidance behavior at the first trial would perhaps spend more time in the inner half of the protected area, and vice versa. These preliminary, exploratory results however did not indicate a significant difference between response types for B6N mice with respect to the average time spent (mean ± SD) in the inner part of the protected area (Cluster A, 19.45 ± 4.59; Cluster B, 22.36 ± 5.37, Wilcoxon-Mann-Whitney $W = 126$, $P = 0.6410$), or the outer part of the protected area (although this latter comparison indicated a trend towards significance; Cluster A, 62.08 ± 9.84; Cluster B, 63.69 ± 6.71, Wilcoxon-Mann-Whitney $W = 90$, $P = 0.0832$). This could indicate that the contrasting patterns of avoidance behavior are indicative for differential adaptive quality of anxiety phenotypes within this strain, but further research is necessary to establish this with more confidence.

Another reason to continue further assessment within each strain is that anxiolytic effects in inbred mice are found to be strain dependent (Clément et al., 2009; Ohl 2003). Whereas diazepam for example reduced anxiety-related behavior in C mice (Ohl et al., 2001b; Belzung and Griebel 2001; Salomons et al., 2012), administration of this compound in B6 mice caused dose-dependent sedative effects in the mHB (Ohl et al., 2001b) and the Elevated Plus Maze (EPM) (Pádua-Reis et al., 2021). In contrast, the use of selective serotoninergic reuptake inhibitors (SSRI's) such as paroxetine (Pádua-Reis et al., 2021) and fluoxetine (Siegmund and Wotjak, 2007) produced anxiolytic effects in B6, whereas C were unresponsive to citalopram – a difference that was attributed to impaired serotoninergic synthesis in C mice (Cervo et al., 2005). Similarly, administration of the metabotropic glutamate receptor 5 (mGlu5R) antagonist MPEP reduced a sensitizing anxiety response in 129P3 mice, but not in C (Salomons et al., 2012). All in all, these examples demonstrate that a correct and more definitive interpretation of the adaptive quality of the individual response types could benefit from further assessment within each strain.

Although such further assessment is essential for an improved understanding of adaptive capacities, it is important to point out that the potential value of these differential response types does not fully depend on the question whether or not the identified response types reflect differential anxiety phenotypes, or whether they reflect individuals that were differentially responsive to the test.

A distinction between responders versus non-responders would be of equal value, as selecting only responsive individuals from an experimental cohort may result in a more relevant animal model (Einat et al., 2018). In addition, one can incorporate this information in the design of animal experiments to improve the quality of a study, which is discussed further on in this chapter.

Genetic influences on inter-individual variability in anxiety

The complex, multidimensional nature of anxiety-related behavior in rodents is also reflected on a genetic level. Anxiety-related behavior in mice has for example been associated with multiple quantitative trait loci (QTLs); the most likely region(s) of a chromosome (or in the mitochondrial DNA) that is/are associated with genetic variation for particular trait (Lander and Botstein, 1989; Clément et al., 2002; Yu et al., 2009). With respect to the expression of unconditioned anxiety, QTL-analyses using different mapping populations of two mouse inbred strains with contrasting anxiety phenotypes (B6 and A) indicate a prominent role for QTLs on chromosomes 1, 10, 15 and 19 (Laarakker et al., 2011, Table 11). For the large part however, these results pertain specifically to the magnitude of the anxiety response, recorded during a single exposure to a stressor. The results of our consomic strain survey on previously collected mHB-data suggests that some of these regions, particularly those located on chromosome 19, may also be associated with habituation of anxiety responses. This finding was in accordance with previous longitudinal assessments in differential mapping populations of B6 and A mice (de Mooij-van Malsen et al., 2013; Gershenfeld and Paul, 1997; Gershenfeld et al., 1997; Zhang et al., 2005). A potential modulating role for chromosomal regions on chromosome 19 was furthermore supported by the fact that significant/suggestive evidence for chromosome 19 remained after controlling for a potentially confounding effect of locomotor activity in the analyses (Chapter 5). While the role of these loci on chromosome 19 in relation to the temporal nature of anxiety responses requires further study, this could indicate that inter-individual variability in adaptive versus non-adaptive responses expressed in the mHB may be genetically determined as well.

Inter-individual variability in anxiety related behavior in rodents has been linked to specific genetic regions or processes that regulate the diversification of this trait within a strain, but reports are scarce. Lathe (2004) argued that sporadic genetic mutation in inbred mice could potentially add to individuality, but that new point mutations are rare and therefore not attributable. He considered it more likely that alterations in so-called hypervariable genetic loci - tandemly repetitive regions (or 'minisatellite variation') play a role in the diversification of behavior (Lathe, 2004). A recent study by Keshavarz et al. (2020) suggests that this may indeed by the case:

natural copy number variation in such a hypervariable genetic locus appeared strongly correlated with consistent individual differences in anxiety-related behavior in B6 mice (Keshavarz et al. 2020). Another recent example comes from a different species, *Drosophila melanogaster*, in which stochastic molecular processes have been found to regulate individual differences in orientation towards visual objects (Takagi and Benton, 2020).

In an attempt to further investigate a genetically determined role for inter-individual variability in habituation of anxiety responses in B6 and A mice, we used the *kml3d* multivariate clustering approach to assess whether a consomic strain survey that was based on individual responses (rather than the group-based comparisons of progenitor strains and consomic lines that were conducted in Chapter 5), would also be suitable procedure for the detection of relevant chromosomal regions. We reasoned that any involvement of a particular chromosome could in theory be reflected in a grouping of mice in which the mice from a consomic line for that chromosome would group together with (a majority of) A mice. Unfortunately application of this clustering procedure on the combined pool of mice that was analyzed in Chapter 5 was not successful, as the gap statistic in all analyses indicated that the partitioning of the data was best reflected by a single cluster.

In these analyses however, the data of both the parental and host strain (B6 and A), as well as the consomic lines were pooled in a single analysis. The absence of any meaningful clusters could potentially again be explained by the previously described risk of the utilized clustering approach: that more subtle effects may end up conflated when pooling different variables/datasets in a single analysis. Namely, further preliminary and exploratory analyses using a univariate *k*-means clustering procedure on a different pool of B6 mice (n = 86, compiled from five different studies) identified two clusters of B6 mice (van Lith, unpublished results): Mice in cluster number 1 displayed significantly higher levels of overall anxiety, avoidance behavior, risk assessment, and significantly lower levels of arousal and exploration than their counterparts in the other cluster (Figure 1). These clusters affected the outcome of a consomic strain survey: with respect to chromosome 19 for example, selecting B6 mice from Cluster Two as a reference group resulted in suggestive evidence for meaningful QTL(s) on chromosome 19, but this evidence disappeared when selecting a reference group of B6 mice from Cluster One (van Lith, unpublished). This does not only demonstrate the fact that the choice for a reference group affects the outcomes of a consomic strain survey, but it also presents potential consequences regarding inter-

individual variability in the consomic lines. After all, if each consomic line harbors 20 chromosomes of the host strain (B6) and one chromosome of the donor strain (A), it is conceivable that any subtypes within the host strain are also somehow reflected in the consomic lines. As said however, these results are preliminary and require further study.

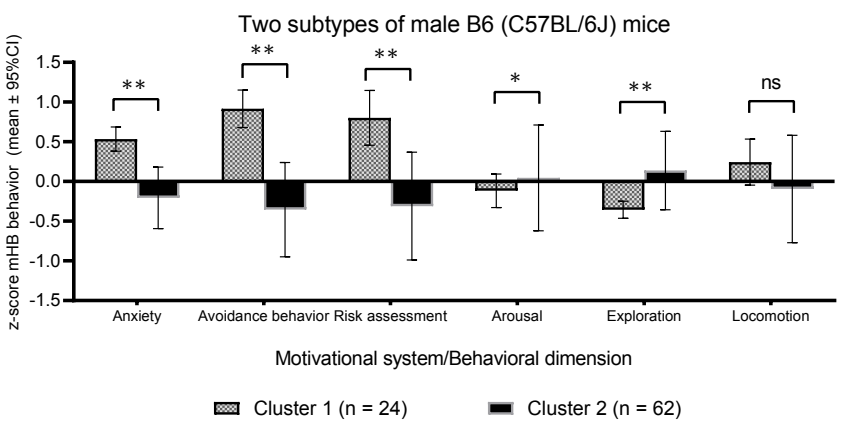


Figure 1. Preliminary results of differences between two clusters (1 and 2) of B6 mice on the motivational system anxiety, and the behavioral dimensions avoidance behavior, risk assessment, arousal, exploration and locomotion. Behavior expressed as integrated behavioral z-scores and presented as means with 95% CI. Effects were significant at **= $P < 0.004$/suggestive at * = $0.004 \leq P < 0.05$ in bootstrapped ANCOVA's with cluster as between-group variable and the ancillary variables 'season' and 'time of day' as covariates.

Inter-individual variability versus the quality of results

A second important contribution of the work presented in this thesis is that it provides several demonstrations of how incorporating individual values in design and analysis improves the quality of laboratory animal research. For one, it demonstrated how considering this variability in the analysis of results may yield new information regarding subtypes of response within and across mouse inbred strains (Chapters 2, 3 and 4). This novel information does not only pertain to the two response types that emerged from the data in these chapters. Defining mice on an individual level also yielded new information regarding the adaptive quality of anxiety responses. An example is the more pronounced non-adaptive anxiety that emerged for 129 mice when re-analyzing the data on an individual level (Chapter 2). In the original experiments of this dataset, impaired

habituation was primarily indicated by an increase in avoidance behavior while any change in (low levels of) activity behavior was less pronounced (Boleij et al. 2012; Salomons et al., 2010a; Salomons et al., 2010b; Salomons et al., 2010c; Salomons et al., 2013). If one considers the approach-avoidance conflicted noted above in this chapter, increased anxiety is assumed to inhibit exploratory activity in the mHB. This was indeed what emerged after zooming in on individual responses: the majority of 129 mice, falling in the non-adaptive anxiety cluster, combined an increase in avoidance behavior with a decrease in exploration (Chapter 2). Another example of novel information that surfaced after zooming in on individual responses is that in Chapters 3 and 4, B6N mice displayed contrasting phenotypes of avoidance behavior: Avoidance behavior increased across trials in some B6N mice, while it decreased in the remaining B6N individuals. These contrasting patterns were not observed on mean strain level, where B6 mice were characterized by a stable pattern of low avoidance behavior in both chapters. This finding demonstrates how latent contrasting phenotypes within an experimental group may mask the detection of an effect on group level, and thereby reduce the power to detect strain-dependent responses (Button et al., 2013; Hager et al., 2014; Lonsdorf and Merz, 2017).

In addition to novel information, this thesis demonstrated how actively accounting for inter-individual variability in the design of an animal experiment improves the experimental design and therefore the quality of an animal experiment (Chapter 4). While the demonstration of confounding effects of inter-individual variability is not new (Barbelivien et al. 2008; Hager et al., 2014; Lonsdorf and Merz, 2017), this experiment to our knowledge is the first that empirically compared the outcomes of a design in which inter-individual variability was accounted for, and a design in which this was not accounted for. Comparing these two designs revealed how inter-individual variability may lead to an overestimation of treatment effects and thereby increase the risk of a Type I error. At the same time it masked confounding influences of other factors, such as experimenter differences which were found for the behavioral dimension avoidance behavior (Chapter 4). The results from this chapter as such provide empirical support for the claim that an active consideration of inter-individual variability in the design of animal experiments benefits the quality, and thereby the reproducibility, of experimental results (Voelkl et al., 2020).

In addition, Chapter 4 provided an example of how such variability could be considered in the design of animal experiments. A central feature in this approach was the a priori characterization of experimental animals. In

addition to the strategy applied in Chapter 4, one could for example also use a priori characterization to ensure that all experimental groups are balanced with respect to any inter-individual variability in a read-out parameter of interest, such that this variability is equally distributed between and within experimental groups (for examples: Labots et al., 2016b; Rojas-Carvajal et al., 2021). The advantage of a priori characterization is that it allows for a systematic incorporation of inter-individual differences in the design of an animal experiment. As such it facilitates better adherence to the two fundamental principles of good design of animal experiments: all variables should be controlled except that due to treatment, ánd treatment and control groups should be identical, with minimal within-group variability (Festing 2014; 2016).

A priori characterization could also be employed to account for any inter-individual variability in a read-out parameter by systematically diversifying this factor in the design of an experiment. This concept, systematic heterogenization, was proposed by Richter (Richter et al., 2010) and entails the systematic incorporation of known sources of variability in the design of animal experiments. The rationale behind this concept is that through systematic inclusion of such factors in a single-lab experiment, this study is presumed to better represent the range of variation one may encounter between studies, which in turn should improve the generalizability of its results (Richter et al., 2010). Sources of variability that have been proposed for this purpose include batch, testing time and experimenter (Paylor, 2009; Bodden et al., 2019; Richter 2020) but also study subjects (i.e. sex, genotype, age, van der Staay, 2006). Systematic heterogenization of individual response profiles could serve as another suitable factor that could be varied across experimental groups. In a randomized complete block design for example, this may be accomplished by matching animals on their response type within each block (as was done in our study), or by balancing the response types within each block. By estimating treatment effects within each block, these should be independent from any block-to-block heterogeneity (Voelkl et al., 2020). Also, if treatment effects are consistent across blocks, these effects are likely more generalizable across studies (Voelkl et al., 2020).

As also discussed in chapter 4 however, we acknowledge that a priori characterization may not be suitable for all phenotypes and contexts. For one, this approach requires some level of consistency of within individuals across time. Although phenotypes such as anxiety and locomotion have been rendered relatively stable over time (Freund et al., 2013; Kazavchinsky

et al., 2019; Keshavarz et al., 2020; Kalueff et al., 2006), other phenotypes (for example grooming) may vary with time within individuals (Kalueff et al., 2006). In addition, some phenotypes may only be expressed during a limited time window (Hofer et al., 2002; Vanderschuren et al., 2016). An alternative approach to adhere to the golden rules of animal experimentation is the use of cross-over designs, in which multiple treatments are contrasted in the same individual, which removes any confounding differences between individuals from the experimental error (Martin and Kraemer, 1987; Bate and Clark, 2014). Cross-over designs however are equally suitable in some, but not all contexts (see discussion chapter 4). In any event, while each of the above described approaches present their strengths and drawbacks, they all facilitate a reduction of within-group (or within-block) variability related to inter-individual differences and thereby increase the statistical power to detect a meaningful effect. In sum, the above results provide several examples of how inter-individual variability may be accounted for in the analysis and design of animal experiments.

Methodological and conceptual limitations

While this thesis contributes to the existing literature by providing an in-depth characterization of inter-individual variability in habituation of anxiety responses, and by providing several examples of how one can account for such inter-individual variability and thereby improve the quality of research, it also presented a number of limitations that are discussed to more detail below.

Temporal stability of anxiety phenotypes

The first constitutes the fact that we did not assess the temporal stability of our identified response types. From a neurobehavioral perspective, the presumed predictive value of inter-individual variability regarding differential susceptibility to treatment, or to development of a particular disorder, only holds if this variability is stable across time (Einat et al., 2018). Also, characterizing experimental animals in a pre-experimental phase with the goal to use this information in experimental design requires some consistency across time of a trait (Chapter 4).

Inter-individual variability in anxiety-related and activity behavior has been rendered stable across time in inbred mice (Ducottet and Belzung, 2004; Kalueff et al., 2006; Lewejohann et al., 2011; Freund et al., 2013; Keshavarz et al., 2020). At the same time however, assessments of such stability are scarce and mostly pertain to B6 mice only. Kalueff et al., (2006) however demonstrated

that exploratory activity was stable across time in 129S1 mice (not tested here), whereas arousal was less consistent. Furthermore, this study suggests that these findings may be generalized to other strains as intercrosses of C and 129S1 as well as NMRI and 129S1 mice displayed the same consistency (Kalueff et al., 2006). Similarly, Ducottet and Belzung (2004) reported temporal consistency in C mice using a chronic mild stress paradigm. The overlap in behavioral profiles between chapters 2, 3 and 4 suggest some temporal consistency for inter-individual variability in anxiety-related behavior for 129S2 and C. Phenotypic variability in natural populations is not regarded as infinite, and often specific traits are correlated with other traits (Koolhaas et al., 2010). This may also hold for the interplay between avoidance behavior and activity that was consistently observed throughout this thesis. Further research is necessary however to establish the temporal consistency of our response types.

Experimenter effects

Second, an unexpected but important finding was the marked experimenter effects we found in both experimental studies (Chapter 3 and 4). On a practical level, the choice for including more than one experimenter was driven by the heavy technical load that came with repeated testing of a substantial number of mice, of multiple inbred strains. As outlined in the section 'Aim of this thesis' (Chapter 1) however, the inclusion of multiple experimenters was also driven by our objective to design our experiments such that it would promote the generalizability of their outcomes. In each of the two studies, two experimenters conducted the behavioral observations, with each experimenter allocated to one of two animal rooms. To control for any confounding influence of this design, both experiments were carefully balanced in the sense that an equal number of mice of all strains were tested by each experimenter, and the test order was randomized across strains for each experimenter (Chapter 3, 4). The inclusion of multiple experimenters in studies on genetic variation in mouse behavior is common, but reports of this factor in data analysis are rare, with only few studies reporting differences in behavioral scores between experimenters (e.g., Bohlen et al., 2014; van Driel and Talling, 2005). According to Bohlen et al. (2014), experimenter effects should not greatly reduce the power to detect treatment effects (i.e. strain differences in habituation of anxiety), provided the experiment was carefully balanced for the inclusion of multiple experimenters, ánd experimenter is included as a factor in the data analysis – which was indeed the case in the experiments reported in Chapter 3, 4 and Box II).

As also outlined throughout this thesis, experimenter effects, but also other factors associated with the laboratory environment, have become increasingly acknowledged as an uncontrollable factor in animal experimentation (Wahlsten et al. 2003; Chesler et al. 2002a; 2002b; Richter 2020). These factors can be roughly divided in those related to 'the human element' (Richter 2020), to the laboratory environment, and experiment-related factors. The following overview discusses some of these factors in relation to the studies in Chapters 3 and 4. Environmental effects related to the 'human element' may include gender-associated olfactory stimuli (i.e. Sorge et al., 2014; de Abreu and Kahluef, 2021), handling experience and familiarity of the animals with the experimenter (i.e. van Driel and Talling, 2005; Gouveia and Hurst, 2017). In both Chapters 3 and 4 the experimenters were of the same sex (female), and the animals were handled by both experimenters from arrival at the test facility onwards. Handling experience differed between experimenters in Chapter 3, but not in Chapter 4, and all experimenters were trained by the same person in handling the mice. Yet, these measures do not preclude the possibility of handling differences – or even differences in other factors such as odour – affecting the outcomes of our study as individual experimenters can "not be subjected to standardization" (Lewejohann et al., 2006).

Next, observer reliability constitutes another experimenter-induced factor when including multiple experimenters and scoring live behavior (Spruijt et al., 2014; Bohlen et al. 2014; Bello and Renter 2018; Richter 2020). In both Chapters 3 and 4, inter-observer reliability was established prior to the start of the study, after a training phase in which both experimenters (per study) aligned their coding by scoring video data from previously collected mHB-studies. As reported, the inter-observer and intra-observer reliability were established at a moderate to good (Chapter 3), and good to excellent level (Chapter 4). Both inter- and intra-reliability however in themselves can be affected by multiple factors, for example behavioral testing experience, the rapidity of the behavior, energy level of the observer etcetera (Bello and Renter 2018; Kaufman and Rosenthal 2009). Especially when conducting behavioral observations over a longer period of time, these factors may induce 'observer-drift': a phenomenon that has been described as 'the implicit changes in code definitions may be observers over time' (Smith, 1986). Unfortunately continuous monitoring of inter- and intra-observer reliability during data collection was not applied in Chapters 3 and 4 and it was not possible to assess this is in retrospect. A potential confounding effect due to observer drift can therefore not be excluded, and applying such continuous monitoring during data collection in

follow up research would be advisable as this would provide more insight in the question whether observer differences played a role. Third, some aspects of the study design in both chapters do not preclude the possibility that experimenter bias may have played a role (Eisenach and Lindner, 2004). The experimenters were for example not blind to the mouse strain they were observing due to different coat colors of the strains (white, C; black, B6N; agouti, 129S2), although in the pharmacological experiment in Chapter 4, the experimenters were blind to treatment (saline/dexmedetomidine).

In addition to these human-element related factors, environmental factors related to the housing facility/test environment ('room effects') form another category that have been shown to affect experimental results (Mogil, 2017). The studies in Chapters 3 and 4 used the same design, in which each experimenter was allocated to one of the animal rooms for the entirety of the study. A result of this design is that any potential animal room effect cannot be dissociated from the found experimenter effects in hindsight. A discussion of the experimenter effects should therefore consider these factors as well. Examples of room effects include ambient temperature (e.g. Pincede et al. 2012), humidity (Chesler et al. 2002a) and light regime (Labots et al. 2016a). These factors were controlled for in both chapters, with no significant differences between ambient temperature and humidity, and the same light regime that was employed in both animal testing rooms. Similarly, husbandry related factors such as diet (Shenk et al. 2020) and schedule of cage change (Pernold et al. 2019) were kept the same between the two rooms in both studies. At the same time, other factors such as barometric pressure (Sato et al. 1999), ambient noise (produced by technical devices, maintenance procedures, conspecifics; Turner et al. 2005) were not specifically controlled for and a potential influence of these factors can therefore not be excluded. A means to dissociate between room effects and experimenter-related effects in future studies could be to have both experimenters observe (the same amount of mice of each strain) in both animal rooms, for example by means of a crossover (e.g. a balanced latin square) design.

Lastly, experimental factors such as testing time, test order, season at test have been shown to affect behavioral phenotyping (Chesler et al. 2002a; Bodden et al. 2019). Testing time and test order was randomized across strains in both chapters and test order did not contribute significantly to the variance in any of the models analyzing the differences between strains.

The above discussion of factors in relation to the results in Chapters 3 and 4 emphasize the importance of carefully balancing environmental factors, such as experimenter, in the design of animal experiments, and accounting for these factors in analysis of the results. What equally remains from this discussion however that is despite careful control for these factors, many other – less easy to standardize – factors may still play a role. The list of factors described above is therefore likely far from exhaustive, with potentially many environmentally induced 'unknown unknowns' that affect experimental results and that we are not even aware of yet (Mogil 2017). A definitive explanation for the found experimenter effects thus can unfortunately not be given. The question then remains how the experimenter effects relate to the identified inter-individual variability (i.e. the response types). It is at this point not possible to completely dissociate the response types from experimenter effects, but the factor experimenter was controlled for by clustering the individual response trajectories on the residual values that remained after controlling for experimenter (and other potentially confounding factors such as test order). This should have taken any variability related to experimenter effects out of the residual variance. As also outlined in various sections of this thesis, automated tracking of behavior in an automated home-cage environment is becoming increasingly suggested as a means to circumvent the confounding influence of a number of the above experimenter-related factors (de Visser et al., 2006; Spruijt et al. 2014; Richter 2020; Voikar and Gaburro, 2020). How the use of automated scoring could contribute to a further evaluation of our response types is discussed in a separate section, further on in this Chapter *("Manual versus automated scoring behavior"*).

Sex differences

Third, in this thesis, all studies were conducted with male mice and this (potentially) limits the impact and conclusions that can be inferred from these studies. Contrary to popular belief, it has been convincingly demonstrated that testing both sexes in factorial designs does not imply a duplication of the total sample size (Shaw et al. 2002; Buch et al. 2019). To our knowledge however, it is unknown whether the same applies to using both sexes in unsupervised cluster analyses. As outlined in Chapter 3 as well as in this chapter, our decision for only using males in the experimental studies was therefore motivated by the fact that our statistical approach of interest – cluster analyses – require a substantial sample size to ensure the detection of meaningful clusters (Dolnicar et al. 2016).

In humans, anxiety disorders are diagnosed twice as frequently in women than in men (Zender and Olshansky 2009). Also, factors such as clinical course and treatment response are known to differ between sexes (Donner and Lowry 2013). Female mice have traditionally been underrepresented in rodent models of psychopathology (Blanchard et al., 1995) – primarily because of the assumption that females show more variability in response due to the estrous cycle (Mogil and Chanda 2005; Prendergast et al. 2014). By means of a meta-analysis Prendergast et al. (2014) however demonstrated that females are no more variable than males. This was for example confirmed in relation to anxiety behavior by Laarakker et al. (2011) who found that females tested at random points in their estrous cycle did not differ in variability from males. Reports of sex differences in mouse models of anxiety are inconsistent, with some studies reporting sex differences, while others found similar patterns between males and females (Armario and Nadal 2013). A detailed overview of literature assessing sex differences in unconditioned, classic tests of anxiety in rodents may be consulted in Donner and Lowry (2013). What follows from this overview is that the studies that do report sex differences indicate lower anxiety levels in female mice compared to males (Donner and Lowry 2013). In addition, general activity in a novel environment is typically found to be higher in females than males (Bothe et al. 2005). With respect to the behavioral phenotype of interest in this thesis, the adaptive quality of anxiety responses, Salomons et al. (2010a) found that female C and 129P3 mice displayed the same contrasting patterns of adaptive (C) versus non-adaptive (129P3) anxiety as their male counterparts. In addition, differential central nervous c-Fos expression was observed between strains in both sexes (Chapter 1, 'Measuring pathological anxiety in mice'). The only differences that were reported between sexes were higher activity levels in female C mice compared to males (in line with previous research, Bothe et al. 2005), and higher post-behavioral test corticosterone levels for female 129P3 compared to their male counterparts (Salomons et al. 2010a).

In line with the above described underrepresentation of females, most studies assessing inter-individual variability in unconditioned anxiety-related behavior in inbred mice used males (e.g Ducottet and Belzung 2004; Jakovcevski et al. 2008; Cohen et al. 2008; Lewejohann et al. 2011). Freund et al. 2013 however found inter-individual variability in exploratory activity that increased with age in female B6N mice. Keshavarz et al. (2020) found no significant differences between male and female B6 mice in anxiety scores, and reported inter-individual variability in anxiety in these mice but did not report to what extent inter-individual variability differed between sexes. Such sex differences in

inter-individual variability have however been assessed in outbred mice and rats. Kazavchinsky et al. (2019) for example reported no sex differences in correlations within individuals that were repeatedly exposed to a forced swim test in CD-1 mice, and individual differences in both male and female CD-1 mice were consistent across time (Kazavchinsky et al. 2019). Pitychoutis et al. (2011) found that responsiveness to an anti-depressant agent was dependent on inter-individual differences and sex in Sprague-Dawley rats, with high novelty (HR) seeking males responding better to treatment than low novelty seeker (LR) males and HR/LR females. Similarly, Carreira et al. (2017) categorized male and female B6 mice as high-responder (HR) and low-responder (LR) phenotypes in the open field test and found sex-dependent effects of the HR/LR phenotype in fear learning but not in tests of unconditioned anxiety (open field, EPM, dark-light box). These studies emphasize the relevance of incorporating sex in studies that address inter-individual differences and an important follow up step of the experimental work in this dissertation should be to generalize these findings to female C, B6N and 129S2 mice.

Analyzing response trajectories

Throughout this thesis, several strategies were used to analyze, summarize and characterize behavioral and physiological response trajectories. The majority of inferential statistics comparing cluster and strain differences were conducted using generalized linear mixed models (GLMMs), which present several advantages over repeated measures ANOVA/ANCOVA, in the sense that they can handle missing values and are more flexible towards non-normally distributed data (Bolker et al., 2009). In addition, these statistical frameworks have been proven effective in accounting for individual-related variability (inter- and intra-variability) through the use of random effects (Bushby et al. 2018).

In Chapter 5 between strain differences were also analyzed using the area under the curve (AUC). This summary metric indicates whether or not a response profile changed over time. A disadvantage of this metric however is that it does not provide information about the directionality of responses. For example, two qualitatively different responses, a rapid increase and a rapid decrease may result in the same AUC value (Reed et al. 2018). The trajectory of a response and its AUC represent two different aspects of temporal responses (Reed et al. 2018). This difference may even affect the outcome of a consomic strain survey, as was shown in Chapter 5, where analysis of response trajectories using GLMM resulted in a suggestive effect for chromosome 19, whereas analysis using the AUC did not. An alternative strategy for summarizing the response trajectories

in Chapter 5 would have been fit each trajectory with a 4th degree polynomial by means of non-linear regression (Motulsky and Ransnas, 1987) and then subsequently use the coefficients of these polynomials to determine their integral (Bailer and Piegorsch, 1990). Lastly, individual-based characterization of behavioral response trajectories was conducted using a clustering procedure specifically designed for multivariate clustering of longitudinal responses: *kml3d*. This procedure is discussed in more detail in the section below.

K-means cluster analysis

As outlined in Chapters 1 and 4, the identification of subtypes of response is not new in the field of behavioral neuroscience, where stratification of subpopulations that display contrasting phenotypes within an experimental pool is a commonly applied technique for investigating the mechanisms underlying any inter-individual variability in behavioral and/or physiological response (Harro 2010; Lonsdorf and Merz 2017). As also outlined in these chapters however, selecting subpopulations based on an artificially predetermined criterion (such as a quantile) is becoming increasingly criticized because such strategies do not fall in line with the conceptualization of human psychopathology as being on a continuum from health to pathology (Insel et al., 2010; Stegman et al., 2019). With ever continuing advancements in data-driven analysis, more data-driven strategies are becoming increasingly advocated (Forkosh, 2021). The presently applied data-driven approach fits this appeal and as such provides a contribution to other applied approaches.

The advantage of unsupervised clustering approaches is that they do not require any normality or parametric assumption within the data (Genolini and Falissard 2010). As such they constitute a category of unbiased, data-driven approaches that can be applied to any kind of data that has some measure of distance between data points (Kiselev et al. 2019). Furthermore they do not require any assumption with respect to the shape of the trajectory, which was particularly suitable for our objective: to classify subgroups of individuals that share a similar response trajectory. This unsupervised nature however also comes with potential challenges, which may constrain the interpretability of the clusters if they are not dealt with appropriately (Frades and Mathiesen, 2010). This section discusses some of the main challenges and how these were handled in this thesis.

a) Selecting the optimal number of clusters

A disadvantage that particularly pertains to *k*-means clustering is that the number of clusters (often denoted as *k*, i.e. the cluster resolution) must be specified by the user prior to analysis (Kiselev et al., 2019). The clustering solution that is returned by *k*-means therefore is the direct result of judgement by the researcher. This poses no problem if one has an a priori idea of the expected number of subgroups (for example through existing literature). Such pre-existing knowledge however is rare, and in most cases *k*-means is employed as an exploratory technique where one has no a priori assumption regarding the cluster resolution (Horne, et al., 2020). This was indeed also the case in this thesis. The study in Chapter 2 was the first time inter-individual variability in habituation of anxiety was assessed in response to the mHB. In Chapter 3, this clustering approach was again employed for the first time while simultaneously clustering individual response trajectories in C, B6N and 129S2 mice. We therefore we did not have any a priori assumption regarding the expected number of clusters on our data. In such cases, it is common to explore the data for multiple number of clusters and evaluate the outcome of each scenario. This was indeed done in our analyses, where we explored the data for clustering resolutions ranging from 2 to 6 clusters (Chapters 2, 3 & 4).

There's an active debate on what constitutes the best method to select the optimal number of clusters (e.g. Kryszczuk and Hurley, 2010; von Luxbourg, 2010; Nies et al., 2019). A common strategy is to use non-parametric quality indices that reflect the relative compactness of within clusters versus the distance between clusters (Genolini and Falissard 2010). A means to enhance the use of quality indices and to optimize accuracy in selecting the optimal number of clusters is to combine multiple quality indices into a single Clustering Validity Index (CVI, Kryszczuk and Hurley 2010; Wahl et al., 2014). To select the number of clusters that best represented our data we therefore used a CVI that combined three of the best validated quality indices: The Calinski-Harabasz criterion (Calinski and Harabasz, 1974), Ray-Turi criteron (Ray and Turi, 1999) and Davies-Bouldin criterion (Davies and Bouldin, 1979).

A second disadvantage inherent to *k*-means is that this algorithm is built around iterative comparisons of differences or distances between clusters, and therefore always minimally assumes two clusters (Horne et al. 2020). *K*-means thus does not assess whether the data is better represented by a single cluster and by default partitions the data into two or more clusters (Frades and Mathiesen, 2010). In order to control for this, the gap statistic was

applied in Chapters 2, 3 & 4 to verify if the data would not be better represented by a single cluster (Tibshirani et al., 2002). The gap statistic compares the within-cluster sum of squares of each clustering solution to a null reference distribution of the data, which then is equivalent to a single cluster and as such gives an indication of whether it is appropriate at all to partition the data into $n > 1$ clusters (Tibshirani et al., 2002). In all chapters, the gap-statistic confirmed that the data possessed enough structure to be partitioned into clusters.

Lastly, as described in Chapter 1, the *k*-means algorithm randomly choses *k* points as initial cluster centers and then iteratively assigns subjects to their closest cluster center until there is a minimal decrease in the squared error (Frades and Mathiesen, 2010). A major drawback of this process is that these initial center points (or seeds) have a strong impact on the final results (Celebi et al., 2013; Kiselev et al. 2019). This bias can be controlled for through repeated application of the *k*-means algorithm (and thus using different random initial selection points) for each *k* number of clusters, using different initial configurations (Genolini et al. 2015). In this thesis we applied this step by using the configuration 'nearlyAll' from the *kml3d*-package (see Genolini et al. 2015 for a detailed description) and for repeating each application of *k*-means 1000 times for each *k* between 2 and 6 clusters, using different initial configuration points (Chapters 2, 3, 4). In all chapters containing cluster analyses, application of the above described steps yielded 2 clusters.

b) Evaluating cluster quality

A next important step in unsupervised clustering is that the quality of the final clusters should be evaluated (Tufféry, 2011). Such evaluation is important, because unsupervised clustering approaches do not infer the likelihood of obtained clusters via statistical tests, i.e. there is no ground truth against which one can formally test the cluster outcome (von Luxburg 2010). In addition, the lack of assumption towards the data makes these techniques sensitive to outliers and/or noise (Tufféry, 2011). A few outlier points can substantially influence the means (centroids) of their respective clusters (Celebi et al. 2013), and random noise may be mistaken for meaningful structure (Celebi et al. 2013; Kiselev et al. 2019). Increasing the sample size may compensate for these risks (Dolnicar et al. 2016).

Small samples sizes may hamper the detection of true number of clusters, or produce clustering solutions that are unstable (Horne et al. 2020). Also, datasets with a low subject to variable ratio may not produce stable clusters. According

to Dolnicar et al. (2016), clustering solutions can be substantially improved by using a sample size from 10 to 30 times the number of variables. Our experiments were designed using a subjects to variable ratio of 30:1 (Chapter 3) and 36:1 (Chapter 4). The study in Chapter 3 originally included 30 extra mice (n = 10/strain) that were not included in the data for Chapter 3. This was the first batch of mice for this study, and this batch was socially housed in groups of $n = 3$ from arrival onwards. Due to prolonged fighting and aggression of group mates in all three strains however, it was ultimately decided to house this batch, and the following batches, individually for the remainder of this experiment. To keep the housing experiences similar between batches, this first batch was removed from further analysis. Due to this unforeseen change in data collection (removal of one batch from analysis, Chapter 3) and unforeseen changes in data analysis (feature selection to improve the cluster quality, see section "*Multidimensionality of individual response types*" and Chapter 4) the ultimately applied ratio was 19.5:1 and 59.7:1 for Chapters 3 and 4 respectively. Following the guidelines by Dolnicar et al. (2016) these ratios were within the suggested range for Chapter 3, and well over the suggested ratio for Chapter 4.

Next, the absence of official likelihood tests emphasize the importance of assessing the stability of the clusters. Cluster stability explores whether the identified structures reflect true patterns in the data, or whether they are due to chance (Frades and Mathiesen, 2010). Cluster stability entails that a cluster should still be identified if the clustering algorithm is run repeatedly on slightly altered datasets (Hennig, 2007). There are various methods to assess cluster stability, for example by randomly dividing the study sample in half and repeat the cluster analysis on each half (Clatworthy et al., 2015). If the clusters are stable, a similar cluster structure should be identified in each half of the sample. Another frequently employed method is bootstrapping with replacement (meaning that a particular individual could occur multiple times in one sample) (von Luxburg, 2010). If repeating the cluster analysis on a large number of bootstrapping samples reveals similar cluster structures as the originally obtained clusters, clusters may be deemed stable (von Luxburg, 2010). Applying bootstrapping with replacement yielded highly stable clusters in Chapters 2 and 4, but not in Chapter 3. As outlined in the Discussion section of Chapter 3, this relative instability may have been related to the aforementioned high impact of the initial center points on the final clustering outcome in relation to the fact that in Chapter 3, response types A and B were close in size. Perhaps the starting points alternated between response type A and B in each bootstrap iteration and this inadvertently contributed to unstable bootstrap results. The

visual representation of the bootstrap analysis for avoidance behavior in Fig. 8a of Chapter 3 indirectly supports this argument: The patterns of avoidance behavior trajectories that 'deviate' from the originally identified response A bare a strong resemblance to the avoidance behavior trajectories from response type B.

It may however also be that the variability in the data between response types simply was less pronounced compared to Chapters 2 and 4, causing individual mice to alternate between response types in the bootstrapping procedure. This was not further explored in Chapter 3, but a means to investigate this could be to turn to a different category of clustering approaches: that of soft or fuzzy clustering (Kiselev et al. 2019). The bootstrapping procedure could categorize individual mice as stable (stably assigned to the same cluster) versus transient (transitioning between clusters) on each iteration and this information could then be used to assign the mice to groups of different probabilities in a fuzzy clustering procedure, rather than distinctive clusters (Peters et al. 2013).

Given the stable bootstrap results however in Chapters 2 and 4, as well as the fact that the behavioral profiles of the identified clusters were similar between Chapters 2, 3, and 4 we are confident that the identified response types likely represent meaningful subtypes of response in the data. In that sense, the employed clustering approach proved an effective method to identify subgroups of mice that follow the same response over time and to delineate their characteristics.

Manual versus automated scoring of behavior

All studies in this thesis were conducted using the mHB, a behavioral test for unconditioned anxiety that is considered especially suitable for multidimensional phenotyping (Ohl et al., 2001b). We selected this paradigm because we did not have any a priori expectations regarding the expression of inter-individual variability in adaptive/non-adaptive anxiety responses when designing the studies. To our knowledge, automated tracking has not yet been validated in the mHB, and has only been applied in semi-automated form, where activity parameters such as velocity, distance traveled, thigmotaxis etc. were scored automatically, while more fine-grained behaviors such as stretched attends (risk assessment), rearing, hole visits/explorations (exploration) and grooming (arousal) were scored manually (Brinks et al. 2007). Given the absence of a priori expectations it was our objective to include these intricate behaviors in the assessment of individual anxiety profiles, and therefore it was decided

to conduct the studies via manual scoring. This allowed us to conduct a fine-grained multidimensional assessment of inter-individuality in habituation of anxiety responses and as such provide a starting point for further assessment and validation of the identified response types.

As discussed in this chapter, the results among others showed that (when concurrently assessing inter-individuality in all five behavioral dimensions) inter-individuality was predominantly expressed in avoidance behavior, exploration and locomotion. Other behaviors, such as risk assessment and arousal did not substantially contribute to the partitioning of the clusters. While further assessment of individuality in risk assessment and arousal could be of interest in further determining inter-individuality in anxiety (risk assessment) or to disentangle inter-individuality in the various forms of grooming (arousal) in these strains, another approach could be to continue further validation of the response types with only the most dominant behavioral dimensions. As also outlined in various sections of this chapter, such further validation would benefit from a more definitive dissociation between anxiety and locomotion, and from establishing temporal stability of these response types. In addition, the found experimenter effects form a potentially confounding factor regarding the interpretation of the identified response types. One interesting strategy of further assessing the adaptive quality of these response types, while accounting for these three factors could be to assess individuality using an automated home-cage.

Automated tracking of behavior in the home-cage environment has indeed become prominently advocated as a way to reduce the incontrollable nature of potentially confounding 'human element' related factors, especially in behavioral studies (de Visser et al., 2006; Spruijt et al. 2014; Richter 2020; Voikar and Gaburro, 2020). This procedure may reduce human interference and bias, and improve time efficiency compared to manual scoring (de Visser et al., 2006; Richter 2020; Voikar and Gaburro 2020). An additional advantage of automated tracking in the home-cage is that behavior can be measured continuously, over hours or even days, instead of minutes as is custom in tests of unconditioned anxiety (Jhuang et al. 2010; Aarts et al. 2015), which enables assessment of the temporal stability of the identified response profiles. Furthermore, automated home-cage environments have been successfully used to dissociate between anxiety-related behavior and locomotion (Kas et al. 2008), and have been successful at measuring anxiety by means of avoidance behavior (the light spot

test, Aarts et al. 2015). These examples demonstrate the potential of automated home cage environments to further assess the adaptive quality of inter-individual variability in anxiety responses.

Other considerations and future prospects

Inter-individual variability in adaptive capacities: Animal welfare.

Assessment of inter-individuality in adaptive and non-adaptive responses may be especially relevant for research concerning animal welfare. Emotional reactivity is a key factor in the domain of welfare-assessment, as understanding the biology of emotions in animals is a prerequisite to safeguard their welfare (Ohl, 2014). Ohl and van der Staay (2012) propose a dynamic concept of animal welfare, which entails that positive welfare is ensured as long as an animal is able to react adequately to both positive and negative stimuli, which enables it to adapt to changing living conditions up to a level it perceives as positive (Ohl and van der Staay, 2012). In this definition, as well as in that of other proponents of dynamic welfare concepts, animal welfare is viewed as a function of adaptation and as such it strongly emphasizes adaptive capacities of animals as an indicator of positive or compromised welfare (e.g., Dantzer and Mormede, 1983; Broom, 1988; Korte et al., 2007; Ohl and van der Staay, 2012).

Most welfare-related studies focusing on the emotional response of animals to a specific stressor measure the magnitude of the (immediate) stress response (Ohl and van der Staay, 2012). The rationale then is that welfare decreases as a stress response increases (Dantzer and Mormede, 1983; Moberg, 1985). Similar to the concept of pathological anxiety outlined in Chapter 1 however, the magnitude of an immediate stress response may simply be an indication of an adaptive response. Therefore, in order to assess which individuals are at risk for compromised welfare, one should also take the temporal nature of a stress response into account (Ohl and van der Staay, 2012). The tools that are used in this thesis to measure adaptive capacities in relation to anxiety in mice could therefore prove highly useful for a temporal based assessment of adaptive capacities in relation to welfare.

Also, they would facilitate individual-based assessment of animal welfare. This is important, as the examples given throughout this thesis demonstrate how individual animals may differ in adaptive capacities in relation to anxiety. It has been demonstrated that differential responses may equally affect the susceptibility of individuals to compromised welfare (Fraser, 2009; Koolhaas and van Reenen, 2016). Ideally thus, such inter-individual variability should be taken into account in welfare assessments (Ortolani and Ohl, 2014; Richter and Hintze, 2019). In

current practice however, welfare guidelines that drive welfare assessments are based on the assumption that some 'universal' welfare exists that is equal for all individuals or a given species under distinct circumstances (Ohl and Putman, 2014; Ortolani and Ohl, 2014; Richter and Hintze, 2019). In these assessments, individual differences in response to environmental stimuli are often being disregarded as statistical noise (Ortolani and Ohl, 2014; Richter and Hinze, 2019). Analogous to developments in preclinical neurobehavioral and bioveterinary research therefore, the importance of considering inter-individual differences when assessing animal welfare is becoming increasingly recognized (Salomons et al., 2009; Richter and Hintze, 2019; Ryan et al. 2019). Within this paradigm shift, the individual-based nature of the tools presented in this thesis could therefore also contribute to a considerable refinement of welfare assessment guidelines.

Recommendations for further study

Various sections of this chapter emphasize further investigation of the identified response types, with the goal to establish whether these response types indeed are indicative of differential adaptive capacities, or whether they merely reflect differential responsivity to the mHB. As discussed in the section "Adaptive quality of individual response types", such evaluation should at least include a more definitive dissociation between anxiety-related behavior and locomotion. This could for example be accomplished by complementing mHB-assessment with tests of unconditioned anxiety that rely less strongly on locomotor activity. One example constitutes the stress-induced hyperthermia paradigm (SIH), which measures the rise in body temperature following a stressful stimulus (see Bouwknecht et al. 2007 for a detailed description). Second, one could assess inter-individual variability in avoidance behavior and activity in behavioral paradigms that are specifically designed to dissociate between anxiety and locomotion, such as the automated home-cage environment that was designed by Kas et al. (2008). In this paradigm, two shelters, one exposed, one not exposed – successfully dissociated anxiety from activity and allowed for longitudinal assessment of behavior in mice (Kas et al., 2008). As recommended earlier, such further evaluation should ideally be conducted within each inbred strain, rather than across strains.

Further evaluation of the adaptive quality of these differential response types could for example be established by means of the three validity requirements that any animal model should fulfil in order to be considered a valid model for a target species: predictive validity, face validity and construct validity (Belzung and Griebel 2001; van der Staay, 2006; van der Staay et al., 2009). Although in

general the target species can either be animal or human, the discussion in the following sections only pertains to mouse models of anxiety in relation to humans as a target species.

Face validity entails that an animal model and humans share commonalities in the behavioral and physiological expression of anxiety (van der Staay, 2006). According to the definition of pathological anxiety in animals employed in this thesis this phenomenon may be defined as "a persistent, uncontrollable, excessive, inappropriate and generalized dysfunctional and aversive emotion, triggering physiological and behavioral responses lacking adaptive value. Pathological anxiety-related behavior in animals is a response to the exaggerated anticipation or perception of threats, which is incommensurate with the actual situation" (Ohl et al., 2008). According to Salomons (2011) the persistent sensitization of anxiety-related behavior, which was displayed by 129 mice in response to the mHB may constitute a form of face validity, as this form of non-adaptive anxiety was displayed in a novel, but not harmful environment. The present thesis suggests that adaptive versus non-adaptive anxiety responses may differ between C, B6N and 129 mice. A more definitive dissociation between anxiety-related behavior and locomotion, as advocated above, would enable a better evaluation of face validity of these differential response types.

Next, predictive validity implies that a therapeutic (i.e. pharmacological) correlation should be demonstrated between the clinical (human) and preclinical (animal) situation (van der Staay, 2006). With respect to mouse models of anxiety in general, anxiolytic and anxiogenic drugs have been observed to produce similar effects on anxiety related behavior in mice and humans (Blanchard et al., 2001; Leonardo and Hen, 2006). Predictive validity in relation to the identified subtypes of response could then be explored by establishing whether the identified subtypes respond differently to pharmacological treatment. A differential response to anxiolytics has primarily been demonstrated in outbred populations. In CD-1 mice for example, administration of the anti-depressant imipramine was only effective in mice that were classified as high-immobile, and not in individuals that were characterized as low-immobile (Vaugeois et al., 1997). Similarly, Pitychoutis et al. (2011) showed how high novelty seeking male Sprague Dawley rats responded better to treatment with an anti-depressant agent than low novelty seeking rats. A few studies indicate that such inter-individual responses to pharmacological compounds may also exist within mouse inbred strains but

reports are scarce. Reddy and Devi (2006) demonstrated that C mice differing in anxiety levels responded differently to the benzodiazepine nitrazepam. Boleij (2013) furthermore found that C could be divided in to responders versus non-responders in a judgement bias test, where responders were characterized by avoidance of a negative stimulus, and non-responders by a lack of avoidance towards this stimulus. Similar to the observation by Reddy and Devi, these two subgroups responded differently to treatment with diazepam (Boleij, 2013). In the present thesis, administration of the alpha 2A-adrenergic receptor agonist dexmedetomidine resulted in a suggestive anxiogenic and marked sedative effect in C, 129S2 and B6N mice, but the two response types did not respond differently to this compound (Chapter 4). Considering the fact that strains may differ in a2-adrenergic receptor binding, administration of strain-dependent doses would have probably been more appropriate (Wilkinson and Manchester, 1983; Fairbanks et al., 2009). In an OF-study comparing C and 129P3 by Salomons et al. (2012), predictive validity was indicated by the fact that a sensitized anxiety response was reversed by administration of the metabotropic glutamate receptor 5 (mGlu5R) antagonist MPEP in 129P3 mice, whereas the benzodiazepine diazepam produced an anxiolytic effect in C. It would be interesting to evaluate whether the two response types also respond differentially to these compounds.

Lastly, construct validity comprises that neurobiological mechanisms underlying anxiety are similar between humans and animals (van der Staay, 2006). With respect to the neurobiological mechanisms that underlie pathological anxiety in humans, it has been suggested that pathological anxiety may be viewed as a form of cognitive dysfunction, leading to impaired integration of information on a higher cognitive level which in turn may lead to inappropriate emotional responses (McNaughton, 1997). In the studies by Salomons et al. (2010a; 2010c) contrasting anxiety-responses between C and 129P3 mice were associated with differential c-Fos expression in brain regions that are involved in higher cognitive processes that regulate the stress-response, thereby providing a first indication of construct validity (Salomons 2011). Such further assessment of any differential c-Fos expression could provide further insights regarding construct validity of our response types.

Summarizing statement and conclusions

Altogether, the work presented in this thesis demonstrates how incorporating individual values in design and analysis may improve the quality of laboratory animal research. Zooming in on individual responses did not only result

in a more comprehensive understanding of inter-individuality in anxiety responses in inbred mice, but it also yielded new information that would be missed when only focusing on average group responses. In addition, we demonstrated how controlling for this type of variability may affect the quality of experimental results. Our results therefore do not only contribute to an improved understanding of potentially differential mechanisms that underlie pathological anxiety in mouse models, but it also contributes to the existing literature that explores novel approaches and perspectives on inter-individual variability in animal experimentation with the goal to improve the quality and reproducibility of experimental results.

References

Aarts, E., Maroteaux, G., Loos, M., Koopmans, B., Kovacevic, J., Smit, A.B., Verhage, M., van der Sluis, S., The Neuro-BSIK Mouse Phenomics Consortium., 2015. The light spot test: Measuring anxiety in mice in an automated home-cage environment. Behav. Brain. Res. 294, 123-150, https://doi.org/10.1016/j.bbr.2015.06.011.

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

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. http://doi.org/ 10.3389/fphar.2013.00137.

Augustsson, H., Meyerson, B.J., 2004. Exploration and risk assessment: a comparative study of male house mice (Mus musculus musculus) and two laboratory strains. Phys. Behav. 81 (4), 685-698, http://doi.org/10.1016/j.physbeh.2004.03.014.

Bailer, A.J., Piegorsch, W.W., 1990. Estimating integrals using quadrature methods with an application in pharmacokinetics. Biometrics 46, 1201-1211, https://doi.org/10.2307/2532462.

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

Barnett, S.A., 1975. The Rat: a study in behavior. University of Chicago Press, Chicago, Il.

Bate S, Clark R., 2014. Experimental Design. In: The Design and Statistical analysis of Animal Experiments, Cambridge UK: Cambridge University Press, pp. 30-121.

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

Belzung, C., Le Pape, G., 1994. Comparison of different behavioral test situations used in psychopharmacology for measurements of anxiety. Physiol. Behav. 56 (3), 623–628, https://doi.org/10.1016/0031-9384(94)90311-5.

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

Beynen, A.C., Gärtner, K., van Zutphen, L.F.M., 2003. Chapter 5: Standardization of animal experimentation. In: Principles of Laboratory Animal Science, Revised edition. Elsevier, Amsterdam.

Blanchard, D.C., Griebel, G., Blanchard, R.J., 1995. Gender bias in the preclinical psychopharmacology of anxiety – Male models for (predominantly) female disorders. J. Psychopharmacol. 9, 79-82, https://doi.org/10.1177/026988119500900201.

Blanchard, D.C., Griebel, G., Blanchard, R.J., 2001. Mouse defensive behaviors: pharmacological and behavioral assays for anxiety and panic. Neurosci. Biobehav. Rev. 25 (3), 205-218, https://doi.org/10.1016/s0149-7634(01)00009-4.

Blanchard, D.C., Griebel, G., Pobbe, R., Blachard, R.J., 2011. Risk assessment as an evolved threat detection and analysis process. Neurosci. Biobehav. Rev. 35, 991-998, https://doi.org/j.neubiorev.2010.10.016.

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

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

Boleij, H., 2013. Emotional perceptions in mice: studies on judgement bias and behavioral habituation. Utrecht University, https://dspace.library.uu.nl/handle/1874/276185.

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, http://doi.org/10.1371/journal.pone.0042544.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, H.H., White, J.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24 (3), 127-135, https://doi.org/10.1016/j.tree.2008.10.008.

Bothe, G.W.M., Bolivar, V.J., Vedder, M.J, Geistfeld, J.G., 2005. Behavioral differences among fourteen inbred mouse strains commonly used as disease models. Comp. Med. 55 (4), 326-334, PMID: 16158908.

Bouwknecht, J.A., Olivier, B., Paylor, R.E., 2007. The stress-induced hyperthermia paradigm as a physiological animal model for anxiety: A review of pharmacological and genetic studies in the mouse. Neursci. Biobehav. Rev. 31, 41-59, https://doi.org/10.1016/j.neubiorev.2006.02.002.

Brinks. V., van der Mark, M., de Kloet, R., Oitzl, M., 2007. Emotion and cognition in high and low stress sensitive mouse strains: a combined neuroendocrine and behavioral study in BALB/c and C57BL/6 mice. Front. Behav. Neurosci. 1 (8), https://doi.org/10.3389/neuro.08.008.2007.

Broom, D.M., 1988. The scientific assessment of animal welfare. Appl. Anim. Behav. Sci. 20 (1-2), 5-19, https://doi.org/10.1016/0168-1591(88)90122-0.

Buch, T., Moos, K., Ferreira, F.M., Fröhlich, H, Gebhard, C., Tresch, A., 2019. Benefits of a factorial design focusing on inclusion of female and male animals in one experiment. J. Mol. Med. 97, 871-877, https://doi.org/10.1007/s00109-019-01774-0.

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.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376 https://doi.org/10.1038/nrn3475.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. Theory Methods, 3(1), 1–27, https://doi.org/10.1080/03610927408827101.

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.

Celebi, M.E., Kingravi, H.A., Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl. 40, 200-210, https://doi.org/10.1016/j.eswa.2012.07.021.

Cervo, L., Canetta, A., Calcagno, E., Burbassi, S., Sacchetti, G., Caccia, S., et al., 2005. Genotype-dependent activity of tryptophan hydroxylase-2 determines the response to citalopram in a mouse model of depression. J. Neurosci. 25, 8165–8172, https://doi.org/10.1523/JNEUROSCI.1816-05.2005.

Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L., Mogil, J.S., 2002a. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. Neurosci. Biobehav. Rev. 26, 907-923, https://doi.org/10.106/s0149-7634(02)00103-3.

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

Clément, Y., Calatayud, F., Belzung, C., 2002. Genetic basis of anxiety-like behaviour: A critical review. Brain. Res. Bull., 57 (1), 57-71, https://doi.org/S0361-9230(0)00637-2.

Clément, Y., Le Guisquet, A., Venault, P., Chapouthier, G., Belzung, C., 2009. Pharmacological alterations of anxious behavior in mice depending on both strain and the behavioral situation. PLoS ONE, 4(11), e7745, https://doi.org/10.1371/journal.pone.0007745.

Cockrem, J. F., 2013. Individual variation in glucocorticoid stress responses in animals. Gen. Comp. Endocrinol. 181, 45-58, https://doi.org/10.1016/j.ygcen.2012.11.025.

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

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

Cryan, J.F., Holmes, A., 2005. The ascent of mouse: advances in modelling human depression and anxiety. Nat. Rev. Drug Discov. 4 (9), 775–790, https://doi.org/10.1038/nrd1825.

Dantzer, R., Mormede, P., 1983. Stress in farm animals: a need for re-evaluation. J. Anim. Sci. 57 (1), 6-18, https://doi.org/10.2527/jas1983.5716.

Davies D.L., Bouldin D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1(2), 224–227, https://doi.org/10.1109/TPAMI.1979.4766909.

De Abreu, M.S., Kalueff, A.V., 2021. Of mice and zebrafish: the impact of the experimenter identity on animal behavior. Lab Anim. 50, 7, https://doi.org/10.1038/s41684-020-00685-9.

De Boer, S.F., Buwalda, B., Koolhaas, J.M., 2016. Untangling the neurobiology of coping styles in rodents: towards neural mechanisms underlying individual differences in disease susceptibility. Neurosci. Biobehav. Rev. 74 (Pt. B), 401-422, https://doi.org/10.1016/j.neubiorev.2016.07.008.

De Mooij-van Malsen, A.J.G., van Lith, H.A., Oppelaar, H., Hendriks, J., de Wit, M., Kostrzewa, E., Breen, G., Collier, D.A., Olivier, B., Kas, M.J., 2009. Interspecies trait genetics reveals association of Adcy8 with mouse avoidance behavior and a human mood disorder. Biol. Psychiatry 66 (12), 1123-1130, https://doi.org/j.biopsych.2009.06.016.

De Visser, L., van den Bos, R., Kuurman, W. W., Kas, M. J. H., Spruijt, B. M., 2006. Novel approach to the behavioural characterization of inbred mice: automated home cage observations. Genes Brain Behav. 5 (6), 458-466, https://doi.org/10.1111/j.1601-183X.2005.00181.x.

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. Physiol., 465, 601-626, https://doi.org/10.1007/s00424-013-1271-7.

Ducottet, C., Belzung, C., 2004. Behaviour in the Elevated-Plus-Maze predicts coping after subchronic mild stress in mice. Physiol. Behav. 81 (3), 417-426, https://doi.org/10.1016/j.physbeh.2004.01.013.

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

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.

Eisenach, J.C., Lindner, M.D., 2004. Did experimenter bias conceal the efficacy of spinal opioids in previous studies with the spinal nerve ligation model of neuropathic pain? Anesthesiology, 100, 765-767, https://doi.org/10.1097/00000542-200404000-00003.

Estanislau, C., 2012. Cues to the usefulness of grooming behavior in the evaluation of anxiety in the elevated plus-maze. Psychol. Neurosci. 5(1), 105-112, https://doi.org/10.3922/j.psns.2012.1.14.

Fairbanks, C.A., Kitto, K.F., Nguyen, H.O., Stone, L.S., Wilcox, G.L., 2009. Clonidine and dexmedetomidine produce antinociceptive synergy in mouse spinal cord. Anesthesiology 110, 648-647, https://doi.org/10.1097/ALN.0b013e318195b51d.

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

Festing, M.F.W., 2016. Study Design. In: Martin-Kehl, MI, Schubiger, PA, editors. Animal Models for Human Cancer: Discovery and Development of Novel Therapeutics. Wiley-VCH, Weinheim; pp. 27-40.

Festing, M.F.W., Baumans, V., Combes, R.D., Halder, M., Hendriksen, C.F.M., Howard, B.R., Lovell, D.P., Moore, G.J., Overend, P., Wilson, M.S., 1998. Reducing the of laboratory animals in biomedical research: problems and possible solutions. Altern. Lab. Anim. 26 (3), 283-301, PMID: 26042346.

Forkosh, O., 2021. Animal behavior and animal personality from a non-human perspective: Getting help from the machine. Patterns, 2, https://doi.org/10.1016/j.patter.2020.100194.

Frades, I., Matthiesen, R., 2010. Overview on techniques in cluster analysis. In: Bioinformatics Methods in Clinical Research. Methods in Molecular Biology (Methods and Protocols), vol F. Humana Press.

Fraser, D., 2009. Animal behavior, animal welfare and the scientific study of affect. Appl. Anim. Behav. Sci. 118, 108-117, https://doi.org/10.1016/j.applanim.2009.02.020.

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, http://doi.org/10.1126/science.1235294.

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/.

Gershenfeld, H.K., Neumann, P.E., Mathis, C., Crawley, J.N., Li, X., Paul, S.M., 1997. Mapping quantitative trait loci for open-field behavior in mice. Behav. Genet., 27 (3), 201-210, https://doi.org/10.1023/a:1025653812535.

Gershenfeld, H. K., Paul, S. M. 1997. Mapping quantitative trait loci for fear-like behaviors in mice. Genomics, 46 (1), 1-8, https://doi.org/10.1006/geno.1997.5002.

Golbidi, S., Frisbee, J.C., Laher, I., 2015. Chronic stress impacts the cardiovascular system: animal models and clinical outcomes. Am. J. Physiol. Heart Circ. Physiol. 308 (12), H1476-H1498, https://doi.org/10.1152/ajpheart.00859.2014.

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

Grewal, S.S., Shepherd, J.K., Bill, D.J., Fletcher, A., Dourish, C.T., 1994. Behavioral and pharmacological characterization of the canopy stretched attend posture test as a model of anxiety in mice and rats. Psychopharmacology, 133, 29-38, https://doi.org/10.1007/s002130050367.

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 behavior and risk assessment of inbred mice. Front. Behav. Neurosci. 8, 314, https://doi.org/10.3389/fn-beh.2014.00314.

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

Henderson, N.D., Turri, M.G., DeFries, J.C., Flint, J., 2004. QTL analysis of multiple behavioral measures of anxiety in mice. Behav Genet. 34(3), 267-293, https://doi.org/10.1023/B:BEGE.0000017872.25069.44.

Hennig, C., 2007. Cluster-wise assessment of cluster stability. Comput. Stat. Data Anal. 52 (1), 258-271, https://doi.org/10.1016/j.csda.2006.11.025.

Herman, J.P., McKlveen, J.M., Ghosal, S., Kopp, B., Wulsin, A., Makinson, R., Scheimann, J., Myers, B., 2016. Regulation of the hypothalamic-pituitary-adrenocortical stress response. Compr. Physiol. 6, 603-621, https://doi.org/10.1002/cphy.c150015.

Hettema, J.M., Neale, M.C., Kendler, K.S., 2001. A review and meta-analysis of the genetic epidemiology of anxiety disorders. Am. J. Psychiatry. 158(10), 1568–1578, https://doi.org/10.1176/appi.ajp.158.10.1568.

Hofer, M.A., Shair, H.N., Brunelli, S.A., 2002. Ultrasonic vocalizations in rat and mouse pups. Curr. Protoc. Neurosci. Chapter 8, Unit 8.14, https://doi.org/10.1002/0471142301.ns0814s17.

Horne, E., Tibble, H., Sheikh, A., Tsanas, A., 2020. Challenges in clustering multimodal data: review of applications in asthma subtyping. JMIR Med. Inform. 8 (5), e164522, https://doi.org/10.2196/16452.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167 (7), 748–751, https://doi.org/10.1176/appi.ajp.2010.09091379.

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

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

Jhuang, H., Garrote, E., Mutch, J., Yu, X., Khilnani, V., Poggio, T., Steele, A.D., Serre, T., 2010. Automated home-cage behavioral phenotyping of mice. Nat. Commun. 1, 68, https://doi.org/10.1038/ncomms1064.

Kakihana, R., Moore, J.A., 1976. Circadian rhythm of corticosterone in mice: The effect of chronic consumption of alcohol. Phsychopharmacologia, 46, 301-305, https://10.1007/BF00421118.

Kalueff, A.V., Tuohimaa, P., 2005a. Mouse grooming microstructure is a reliable anxiety marker bidirectionally sensitive to GABAergic drugs. Eur J Pharmacol. 508(1-3), 147-153, https://doi.org/10.1016/j.ejphar.2004.11.054.

Kalueff, A.V., Tuohimaa, P., 2005b. Contrasting grooming phenotypes in three mouse strains markedly different in anxiety and activity (129S1, BALB/c and NMRI). Behav. Brain Res. 160, 1-10, https://doi.org/10.1016/j.bbr.2004.11.010.

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

Kas, M.J.H., de Mooij-van Malsen, J.G., Olivier, B., Spruijt, B.M., van Ree, J.M., 2008. Differential genetic regulation of motor activity and anxiety-related behaviors in mice using an automated home cage task. Behav. Neurosci. 122 (4), 769-776, https://doi.org/10.1037/0735-7044.122.4.769.

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

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/10.1016/j.vascn.2018.11.007.

Keshavarz, M., Krebs-Watson, R., Refki, P., Savriama, Y., Zhang, Y., Guenther, A., Brückl, T. M., Binder, E. B., Tautz, D., 2020. Natural copy number variation differences of tandemly repeated small nucleolar RNAs in the Prader-Willi syndrome genomic region regulate individual behavioral responses in mammals. bioRxiv 476010, https://doi.org/10.1101/476010.

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

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, http://doi.org/10.1016/j.yfrne.2010.04.001.

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.

Korte, S.M., 2001. Corticosteroids in relation to fear, anxiety and psychopathology. Neurosci. Biobehav. Rev. 25 (2), 117-142, https://doi.org10.1016/S0149-7634(01)00002-1.

Korte, S.M., Olivier, B., Koolhaas, J.M., 2007. A new animal welfare concept based on allostasis. Physiol. Behav. 92 (3), 422–428, https://doi.org/10.1016/j.physbeh.2006.10.018.

Kryszczuk, K., Hurley, P., 2010. Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. In: El Gayar, N., Kittler, J., Roli, F. (eds). Multiple Classifier Systems. MCS 2010. Lecture Notes in Computer Science, vol 5997. Springer, Berlin, Heidelberg. 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, http://doi.org/10.1016/j.physbeh.2010.10.019.

Labots, M., Zheng, X., Moattari, G., Ohl, F., van Lith, H.A., 2016a. Effects of light regime and substrain on behavioral profiles of male C57BL/6 mice in three tests of unconditioned anxiety. J. Neurogenet. 30 (3-4), 306-315, https://doi.org/10.1080/01677063.2016.1249868.

Labots, M., Zheng, X., Moattari, G., Lozeman-van 't Klooster, J.G., Baars, J.M., Hesseling, P., Lavrijsen, M., Kirchhoff, S., Ohl, F., van Lith, H.A., 2016b. Substrain and light regime effects on integrated anxiety-related behavioral z-scores in male C57BL/6 mice – Hypomagnesaemia has only a small effect on avoidance behavior. Behav. Brain Res. 306, 71-83, https://doi.org/10.1016/j.bbr.2016.01.060.

Lander, E. S., Botstein, D., 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121 (1), 186-199, PMCID: PMC1203601.

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

Leonardo, E.D., Hen, R., 2006. Genetics of affective and anxiety disorders. Annu. Rev. Psychol. 57, 117-137, https://doi.org/10.1146/annurev.psych.57.102904.190118.

Lepicard, E.M., Joubert, C., Hagneau, I., Perez-Diaz, F., Chapouthier, G., 2000. Differences in anxiety-related behavior and responses to diazepam in BALB/cByJ and C57BL/6J strains of mice. Pharmacol. Biochem. Behav. 67 (4), 739-748, https://doi.org/10.1016/S0091-3057(00)00419-6.

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

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

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

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.

Martin, P., Kraemer, H.C., 1987. Individual differences in behavior and their statistical consequences. Anim. Behav. 35 (5), 1366-1375, https://doi.org/10.1016/S0003-3472(87)80009-X.

McNaughton, N., 1997. Cognitive dysfunction resulting from hippocampal hyperactivity – A possible cause of anxiety disorder? Pharmacol. Biochem. Behav. 56, 603-611, https://doi.org/10.1016/s0091-3057(96)00419-4.

Menard, C., Pfau, M.L., Hodes, G.E., Russo, S.J., 2016. Immune and neuroendocrine mechanisms of stress vulnerability and resilience. Neuropsychopharmacology, 42 (1), 62-80, https://doi.org/10.1038/npp.2016.90.

Merikangas, K.R., Pine, D., 2002. American College of Neuropsychopharmacology. Genetic and Other Vulnerability Factors for Anxiety and Stress Disorders. In: Neuropsychopharmacology: The Fifth Generation of Progress. Lippincott, Williams & Wilkins; Nashville, TN, USA, p. 867.

Milner, L.C., Crabbe, J.C., 2008. Three murine anxiety models: results from multiple inbred strain comparisons. Genes Brain Behav. 7(4), 496-505, https://doi.org/10.1111/j.1601-183X.2007.00385.x.

Moberg, G.P., 1985. Biological response to stress. In: G.P. Moberg (Ed.), Animal Stress, American Physiological Society, Bethesda, Maryland, pp. 27-49.

Mogil, J.S., 2017. Laboratory environmental factors and pain behavior: the relevance of unknown unknowns to reproducbility and translation. Lab Anim. 46 (4), 136-141, https://doi.org/10.1038/laban.1223.

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.

Motulsky, H.J., Ransnas, L.A., 1987. Fitting curves to data using nonlinar regression: a practical and nonmathematical review. FASEB J. 1 (5), 365-374, https://doi.org/10.1096/fasebj.1.5.3315805.

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

Nies, H.W., Zakaria, Z., Mohamad, M.S., Chan, W.H., Zaki, N., Sinnott, R.O., Napis, S., Chamoso, P., Omatu, S., Corchado, J.M., 2019. A review of computational methods for clustering genes with similar biological functions. Processes, 7 (9), 550, https://doi.org/10.3390/pr7090550.

Ohl, F., 2003. Testing for anxiety. Clin. Neurosci. Res. 3 (4-5), 233-238, https://doi.org/10.1016/S1566-2772(03)00084-7.

Ohl, F., 2005. Animal models of anxiety. Handb Exp Pharmacol. 169, 35-69, https://doi.org/10.1007/3-540-28082-0_2.

Ohl, F., 2014. Welfare phenotypes in mice: on the biological value of individual variation of self-perception - PhD project proposal. Programme Emotion & Cognition, Department of Animals in Science and Socety, Faculty of Veterinary Medicine, Utrecht University.

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

Ohl, F., Sillaber, I., Binder, E., Keck, M.E., Holsboer, F., 2001b. Differential analysis of behavior and diazepam-induced alterations in C57BL/6N and BALB/c mice using the modified hole board test. J. Psychiatr. Res. 35 (3), 147-154, https://doi.org/10.1016/S0022-3956(01)00017-6.

Ohl, F., Roedel, A., Binder, E., Holsboer, F., 2003. Impact of hight and low anxiety on cognitive performance in a modified hole board test in C57BL/6 and DBA/2 mice. Eur. J. Neurosci. 17, 128-136, https://doi.org/10.1046/j.1460-9568.2003.02436.x.

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.

Ohl, F., van der Staay, F.J., 2012. Animal welfare: At the interface between science and society. Vet. J. 192 (1), 13-19, https://doi.org/10.1016/j.tvjl.2011.05.019.

Ohl, F., Putman, R.J., 2014. Animal welfare at the group level: More than the sum of individual welfare? Acta Biotheor. 62, 35-45, https://doi.org/10.1007/s10441-013-9205-5.

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.

Ortolani, A., Ohl, F., 2014. Hondenwelzijn: een nieuw perspectief. Utrecht: Universiteit Utrecht.

Pádua-Reis, M., Nôga, D.A., Tort, A.B.L., Blunder, M., 2021. Diazepam causes sedative rather than axiolytic effects in C57BL/6 mice. Sci. Rep. 11, 9335, https://doi.org/10.1038/s41598-021-88599-5.

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

Pernold, K., Ianello, F., Low, B.E., Rigamonti, M., Rosati, G., Scavizzi, F., Wang, J., Raspa M., Wiles, M.V., Ulfhake, B., 2019. Towards large scale automated cage monitoring – diurnal rhythm and impact of interventions on in-cage activity of C57BL/6J mice recorded 24/7 with a non-disrupting capacitive-based technique. PLoS ONE 14, e0211063, https://doi.org/10.1371/journal.pone.0211063.

Peters, G., Crespo, F., Lingras, P., Weber, R., 2013. Soft clustering – fuzzy and rough approaches and their extensions and derivatives. Int. J. Approx. Reason. 54, 307-322, https://doi.org/10.1016/j.ijar.2012.10.003.

Pincede, I., Pollin, B., Meert, T., Plaghki, L., Le Bars, D., 2012. Psychophysics of a nociceptive test in the mouse: ambient temperature as a key factor for variation. PLoS ONE 7, e36699, https://doi.org/10.1371/journal.pone.0036699.

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.

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

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.

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.

Ray, S., Turi, R.H., 1999. Determination of number of clusters in *K*-means clustering and application in colour image segmentation. In: Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 137–143.

Reddy, P.V., Devi, K., 2006. Intrastrain variations in anxiolytic effect of nitrazepam in mice. Indian J. Physiol. Pharmacol. 50 (3), 309-312, PMID: 17193905.

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. Lab Animal (NY) 46 (9), 343-349, https://doi.org/10.1038/laban.1330.

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

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

Richter, S.H., Hintze, S., 2019. From the individual to the population – and back again? Emphasizing the role of the individual in animal welfare science. Appl. Anim. Behav. Sci. 212, 1-8, https://doi.org/10.1016/j.applanim.2018.12.012.

Rodgers, R.J., Cole, J.C., 1994. Anxiolytic-like effect of (S)-WAY 100135, a 5-HT_{1A} receptor antagonist, in the murine elevated plus-maze test. Eur. J. Pharmacol. 261, 321-325, https://doi.org/10.1016/0014-2999(94)90124-4.

Rodgers, R., Haller, J., Holmes, A., Halasz, J., Walton, T., Brain, P., 1999. Corticosterone response to the plus-maze: high correlation with risk assessment in rats and mice. Physiol. Behav. 68 (1), 47-53, https://doi.org/10.1016/S0031-9384(99)00140-7.

Rojas-Carvajal, M., Quesada-Yamasaki, D., Brenes, J.C., 2021. The cage test as an easy way to screen and evaluate spontaneous activity in preclinical neuroscience studies. Methodsx 8, 101271, https://doi.org/10.1016/j.mex.2021.101271.

Rougé-Pont, F., Deroche, V., Le Moal, M., Piazza, P. V., 1998. Individual differences in stress-induced dopamine release in the nucleus accumbens are influenced by corticosterone. Eur. J. Neurosci. 10 (12), 3903-3907, https://doi.org/10.1046/j.1460-9568.1998.00438.x.

Ryan, S., Bacon, H., Endenburg, N., Hazel, S., Jouppi, R., Lee, N., Seksel, K., Takashima, G., 2019. WSAVA Animal welfare guidelines for companion animal practitioners and veterinary teams. J. Small Anim. Pract. 60 (5), E1-E46, https://doi.org/10.1111/jsap.12998.

Salomons, A.R., 2011. The anxious mouse: Implications for preclinical research and animal welfare. Utrecht University, https://dspace.library.uu.nl/handle/1874/192214.

Salomons, A.R., Arndt, S.S., Ohl, F. 2009. Anxiety in relation to animal environment and welfare. Scand. J. Lab. Animal. Sci. 36 (1), 37-45, ISSN: 09013393.

Salomons, A.R., Bronkers, G., Kirchhoff, S., Arndt, S.S., Ohl, F., 2010a. Behavioral 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, http://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, http://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, http://doi.org/10.1111/j.1601-183X.2009.00527.x

Salomons, A.R., Espititia Pinzon, N., Boleij, H., Kirchhoff, S., Arndt, S.S., Nordquist, R.E., Lindemann, L., Jaeschke, J., Spooren, W., Ohl, F., 2012. Differential effects of diazepam and MPEP on habituation and neuro-behavioral processes in inbred mice. Behav. Brain Funct. 8, 30, https://doi.org/10.1186/1744-9081-8-30.

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, http://doi.org/10.1016/j.bbr.2013.02.023.

Sato, J., Morimae, H., Seino, Y., Kobayashi, T., Suzuki, N., Mizumura, K., 1999. Lowering barometric pressure aggravates a mechanical allodynia and hyperalgesia in a rat model of neurpathic pain. Neurosci. Lett. 266 (1), 21-24, https://doi.org/10.1016/s0304-3940(99)00260-8.

Sgoifo, A., de Boer, S. F., Haller, J., Koolhaas, J. M., 1996. Individual differences in plasma catecholamine and corticosterone stress responses of wild-type rats. Physiol. Behav. 60 (6), 1403-1407, https://doi.org/10.1016/S0031-9384(96)00229-6.

Shaw, R., Festing, M.F.W., Peers, I., Furlong, L., 2002. Use of factorial designs to optimize animal experiments and reduce animal use. ILAR Journal 43(4), 223-232, https://doi.org/10.1093/ilar.43.4.223.

Shenk, J., Lohkamp, K.J., Wiesmann, M., Kiliaan, A.J., 2020. Automated analysis of stroke mouse trajectory data with traja. Front. Neurosci. 14, 158, https://doi.org/10.3389/fnins.2020.00518.

Siegmund, A., Wotjak, C.T., 2007. A mouse model of posttraumatic stress disorder that distinguishes between conditioned and sensitised fear. J. Psychiatr. Res. 41, 848-860, https://doi.org/10.1016/j.jpsychires.2006.07.017.

Smith, G.A., 1986. Observer drift: A drifiting definition. Behav. Anal. 9 (1), 127-128, https://doi.org/10.1007/BF03391937.

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

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. Meth. 234, 2-12, https://doi.org/10.1016/jneumeth.2014.03.001.

Stegman, Y., Schiele, M.A., Schümann, D., Lonsdorf, T.B., Zwanzger, P., Romanos, M., Reif, A., Domschke, K., Deckert, J., Gamer, M., Pauli, P., 2019. Individual differences in human fear generalization – pattern identification and implications for anxiety disorders. Transl. Psychiatry, 9, 307, https://doi.org/10.1038/s41398-019-0646-8.

Takagi, S., Benton, R., 2020. Animal Behavior: A neural basis of individuality. Sci. Dir. 30 (12), R710-R712, https://doi.org/10.1016/j.cub.2020.04.052.

Tam, W. Y., Cheung, K-K., 2020. Phenotypic characteristics of commonly used mouse inbred strains. J. Mol. Med. Jul 25. https://doi.org/10.1007/s00109-020--1953-4.

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

Tufféry, S., 2011. Data mining and statistics for decision making. Chpt. 9: Cluster analysis. John Wiley & Sons.

Turner, J.G., Parrish, J.L., Hughes, L.F., Toth, L.A., Caspary, D.M., 2005. Hearing in laboratory animals: strain differences and nonauditory effects of noise. Comp. Med. 55 (1), 12-23, PMID: 15766204.

Turri, M.G., Datta, S.R., DeFries, J., Henderson, N.D., Flint, J., 2001. QTL analysis identifies multiple behavioral dimensions in ethological tests of anxiety in laboratory mice. Curr Biol. 11(10), 725-734, https://doi.org/10.1016/s0960-9822(01)00206-8.

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

Van der Staay, F.J., 2006. Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. Brain Res. Rev. 52, 131-159, https://doi.org/10.1016/j.brainresrev.2006.01.006.

Van der Staay, F.J., Arndt, S.S., Nordquist, R.E. 2009. Evaluation of animal models of neurobehavioral disorders. Behav. Brain Funct. 5, 11, https://doi.org/10.1186/1744-9081-5-11.

Vanderschuren, L.J.M.J., Achterberg, E.J.M., Trezza, V., 2016. The neurobiology of social play and its rewarding value in rats. Neurosci. Biobehav. Rev. 70, 86-105, https://doi.org/10.1016/j.neubiorev.2016.07.025.

Vaugeois, J.M., Passera, G., Zuccaro, F., Costentin, J., 1997. Individual differences in resonse to imipramine in the mouse tail suspension test. Psychopharmacology 134 (4), 387-391, https://doi.org/10.1007/s002130050475.

Voelkl, B., Altman, N.S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N.A., Kas, M.J., Schielzeth, H., Van de Casteele, T., Würbel, H., 2020. Reproducibility of animal research in the light of biological variation. Nat. Rev. Neurosci. 2020; 21, 384-393, https://doi.org/10.1038/s41583-020-0313-3.

Voikar, V., Gaburro, S., 2020. Three pillars of automated home-cage phenotyping of mice: Novel findings, refinement, and reproducibility based on literature and experience. Front. Behav. Neurosci. 14, 575434, https://doi.org/10.3389/fnbeh.2020.575434.

Von Luxburg, U., 2010. Clustering stability: an overview. Found. Trends Mach. Learn. 2 (3), 235-274, https://doi.org/10.1561/2200000008.

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

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

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

Young, E.A. Abelson, J.L. Liberzon, I., 2008. Stress Hormones and Anxiety Disorders. In: Handbook of Anxiety and Fear. Academic Press, Oxford, p. 455-473.

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.

Zhang, S., Lou, Y., Amstein, T.M., Anyango, M., Mohibullah, N., Osoti, A., Stancliffe, D., King, R., Iraqi, F., Gershenfeld, H.K., 2005. Fine mapping of a major locus on chromosome 10 for exploratory and fear-like behavior in mice. Mamm. Genome 16 (5), 306-318, https://doi.org/10.1007/s00335-00402427-8.

Nederlandse samenvatting

Van nummer naar individu. Het verbeteren van de kwaliteit van (proef) dierstudies naar (experimenteel) gedrag door het in kaart brengen en meenemen van inter-individuele variatie in de laboratoriummuis.

Verreweg het meeste dierexperimenteel onderzoek naar de biologische grondslag van psychiatrische aandoeningen wordt uitgevoerd aan de hand van muis- en ratmodellen. De muis heeft daarbij de laatste decennia de rat ingehaald qua populariteit als diermodel. Een belangrijke factor hierin is dat immer verbeterende genetische technieken onderzoekers steeds beter in staat stellen om genetische factoren die een rol spelen bij het ontstaan van psychiatrische aandoeningen in kaart te brengen en mee te nemen in hun diermodel. Meer dan 90% van het menselijk genoom is geconserveerd in dat van laboratoriummuizen, wat deze proefdieren uitermate geschikt maakt voor onderzoek naar genetische factoren die een rol spelen bij deze aandoeningen.

Binnen dit type onderzoek heeft men toenemende interesse in de mate waarin individuele muizen van elkaar verschillen in fenotype, zoals gedrag of fysiologie. Onderzoek wijst uit dat deze zogeheten inter-individuele variatie terug te voeren is op een breed scala aan factoren, variërend van (epi-) genetische bronnen en omgevingsfactoren, tot factoren die zijn gekoppeld aan het microbioom. Het ontstaan van inter-individuele variatie bij proefdieren is daarmee het resultaat van complexe interacties die maken dat gedrag of fysiologie subtiel kan verschillen tussen individuen. Binnen diermodellen voor psychiatrische aandoeningen is deze inter-individuele variatie relevant omdat bij de mens qua onstaansgeschiedenis en behandeling van psychiatrische aandoeningen een vergelijkbaar fenomeen speelt: Ook hier zorgen complexe interacties tussen (epi-) genetische en omgevingsbepaalde factoren ervoor dat de ontvankelijkheid voor het ontwikkelen van een psychiatrische stoornis en de effectiviteit van een behandeling, verschilt van persoon tot persoon. Meer inzicht in de biologische grondslag van deze inter-individuele variatie bij muismodellen kan bijdragen aan meer kennis over het ontstaan van psychiatrische aandoeningen bij de mens, en aan het beter afstemmen van een eventuele behandeling op het individu.

De toegenomen interesse in inter-individuele variatie bij proefdieren heeft echter ook nog een andere reden. Binnen dierexperimenteel onderzoek is het gebruikelijk om zowel omgevingsfactoren die van invloed zijn op testresultaten, als de genetische achtergrond van de dieren, zoveel mogelijk op voorhand te benoemen, en (waar mogelijk) gelijk te houden: ook wel

standaardisatie genoemd. De invloed van genetische factoren op de uitkomst van een dierproef kan bijvoorbeeld onder controle worden gehouden door het gebruik van muizen met eenzelfde genetische achtergrond: inteeltmuizen. Deze muis-inteeltstammen worden gebruikt in ongeveer 80% van de muisstudies wereldwijd. De gedachte achter het standardaardiseren van dierexperimenten is dat eventuele gevonden effecten van een test of behandeling aan de toegepaste interventie kunnen worden toegeschreven, en niet aan andere zaken zoals bijv. een verschil in huisvesting of de invloed van genen. De complexiteit van factoren die van invloed zijn op verschillen tussen individuele dieren maakt echter dat zelfs inteeltmuizen binnen een test, ondanks strenge maatregelen van standaardisatie, alsnog verschillend kunnen reageren op dezelfde omstandigheden of testsituatie.

Deze inter-individuele variatie is van invloed op de kwaliteit van proefresultaten, omdat ze een bepaalde ruis introduceert die het lastiger maakt een relevant effect van een interventie te detecteren. Ook maakt deze variatie het lastiger om onderzoeksresultaten tussen verschillende experimenten, en tussen verschillende instituten, te vergelijken en te reproduceren. Dit fenomeen komt in veel verschillende onderzoeksvelden voor, maar krijgt binnen het proefdieronderzoek nog een extra dimensie door de ethische bezwaren die kleven aan het onnodig gebruik van proefdieren. Een veel gebruikte manier om met deze zogenaamde ‘storende variatie’ om te gaan is het verhogen van het aantal proefdieren binnen een experiment. Deze vorm van compensatie heeft tot nu toe echter niet geleid tot een verbeterde vergelijkbaarheid van onderzoeksresultaten. Ook groeit de onderkenning dat – juist door de complexiteit van factoren die een rol spelen bij het tot uiting komen van inter-individuele variatie – het onwaarschijnlijk is dat men door nog verder gaande standaardisatie ooit volledig grip op inter-individuele variatie zal krijgen. De laatste jaren wordt daarom in toenemende mate gepleit voor een andere kijk op inter-individuele variatie: in plaats van het te beschouwen als ongewenste ruis kan deze variatie wellicht juist bijdragen aan zowel de kwaliteit van dierproeven als de kennis die daaruit voortvloeit. Dit proefschrift past binnen deze andere kijk en onderzoekt hoe inter-individuele variatie in gedrag en fysiologie in kaart gebracht kan worden, en in hoeverre deze variatie ingezet kan worden om de kwaliteit van dierproeven te verhogen. Daarbij richten we ons op het meenemen van inter-individuele variatie in de statistische analyse van onderzoeksresultaten, en in het ontwerp van een dierproef.

Deze doelstellingen zijn uitgewerkt aan de hand van een specifiek fenotype: de uiting van angst-gerelateerd gedrag bij muizen. Angststoornissen bij de mens vormen een van de meest voorkomende psychiatrische aandoeningen: Ongeveer 30% van de bevolking krijgt op enig moment in zijn of haar leven te maken met een angststoornis. Angst wordt in principe gezien als een biologisch zinvolle reactie op potentieel gevaar, welke mens en dier uiteindelijk helpt te overleven. Evolutionair gezien is angst dus een zeer nuttige emotie. Pas wanneer een angstreactie langdurig aanhoudt, oncontroleerbaar wordt en extreme vormen aanneemt spreekt men van pathologische angst, wat kan uitmonden in een angststoornis. Van angststoornissen bij de mens is bekend dat ze deels genetisch bepaald zijn, maar ook dat omgevingsfactoren zoals negatieve levenservaringen of stress in de vroege kindertijd een rol spelen. Analoog aan het hierboven beschreven beeld van psychiatrische aandoeningen in het algemeen, werken deze genetische en omgevingsfactoren op elkaar in waardoor de ene persoon onder dezelfde omstandigheden een angststoornis kan ontwikkelen, en de ander niet.

Angst wordt bij muizen hoofdzakelijk gemeten aan de hand van gedragstesten. Een aantal van deze testen meten spontane, aangeboren (ongeconditioneerde) angst. Typerend voor dit soort testen is dat een muis in een nieuwe omgeving wordt geplaatst, welke vervolgens door het dier verkend kan worden. Muizen hebben als prooidier van nature de neiging om open, onbeschermde oppervlakten te vermijden en een nieuwe omgeving te verkennen vanuit de meer beschermde rand of periferie. Dit soort gedragstesten roept bij muizen dan ook een conflict op tussen de drang om een nieuwe omgeving te verkennen, en de neiging om potentieel gevaarlijke situaties te vermijden. Angst-gerelateerd gedrag wordt hier onder andere gemeten als de mate waarin het dier het centrale deel van de testomgeving vermijdt (angstig), dan wel verkent (niet angstig). De uiting van angst bij muizen is echter complex, waarbij niet alleen angst-gerelateerd gedrag wordt vertoond, maar ook verandering in exploratiegedrag en algemeen activiteitsniveau kan worden gemeten. Om een zo volledig mogelijk beeld te verkrijgen van inter-individuele variatie in angst-gerelateerd gedrag worden idealiter deze andere gedragsdimensies ook meegenomen. In dit proefschrift is daarvoor gebruik gemaakt van het modified Hole Board (mHB), een gedragstest die bij uitstek geschikt is voor zo'n multidimensionele aanpak. In **Hoofdstuk 1** wordt deze test nader toegelicht en beschreven.

Een kenmerk van pathologische angst is verder dat het een angstreactie betreft die langdurig aanhoudt en/of oncontroleerbaar is. Zoals hierboven beschreven kan een angstrespons an sich worden gezien als een biologisch zinvolle respons op een potentieel bedreigende situatie. Een eventuele angstreactie na een eerste blootstelling aan een gedragstest kan dus worden gezien als een adaptieve, adequate reactie. Om te kunnen bepalen in hoeverre de angst oncontroleerbaar is of langdurig aanhoudt (met andere woorden om iets te kunnen zeggen over de mate van pathologische angst) is het echter belangrijk om het verloop van een angstrespons in kaart te brengen, in plaats van enkel de directe reactie aan een eerste blootstelling. In dit proefschrift zijn de muizen daarom meerdere malen getest in het mHB om het verloop van de angstrespons per muis te kunnen meten. De gedachtegang daarbij was dat men bij een adaptieve angstrespons zou mogen verwachten dat deze afneemt na herhaaldelijke blootstelling aan de test. Bij een niet-adaptieve angstrespons neemt de angstreactie niet af na herhaaldeliijke blootstelling, of zelfs toe. In het eerste geval is sprake van habituatie van de angstrespons, en in het tweede geval van sensitisatie.

Het verloop van de angstrespons werd in dit proefschrift gemeten in drie verschillende muis-inteeltstammen: C57BL/6, BALB/c en 129. Deze stammen behoren tot de meest gebruikte muis-inteeltstammen in biomedisch proefdieronderzoek en staan bekend om hun verschillen in basaal angstniveau. Aan de hand van cluster analyse werd vervolgens onderzocht in hoeverre individuen binnen dezelfde stam, of tussen stammen, systematisch samen groeperen over meerdere gedragsdimensies, en op fysiologisch vlak: met andere woorden, of wij individuele, multidimensionele responstypen konden onderscheiden in onze gegevens, hoe deze profielen eruit zouden zien en hoe deze zouden zijn verspreid tussen en binnen verschillende muis-inteeltstammen.

Als eerste verkenning maakten we in **Hoofdstuk 2** daarbij gebruik van reeds bestaande data om te bepalen in hoeverre inteeltmuizen inderdaad systematisch samen groeperen over meerdere gedragsdimensies. De gegevens van deze dataset kwam van een serie reeds uitgevoerde experimenten waarbij de mate waarin BALB/c muizen en 129 muizen konden wennen aan herhaaldelijke blootstelling aan het mHB werd onderzocht. In de oorspronkelijke studies kwam steevast naar voren dat BALB/c muizen een sterke angstreactie laten zien bij een eerste blootstelling aan het mHB, maar dat deze angstrespons rap afneemt bij herhaaldelijke blootstelling, terwijl exploratiegedrag en

algemene activiteit toenamen. Dit gedragsprofiel werd gekwalificeerd als een adaptieve angstreactie. In tegenstelling tot BALB/c muizen vertoonden 129 muizen weinig angst bij een eerste blootstelling aan het mHB, maar nam het angstgedrag toe naarmate de dieren vaker werden getest in het mHB. Het profiel van 129 muizen werd daarbij gekwalificeerd als niet-adaptieve angst. In **Hoofdstuk 2** analyseerden wij de gegevens van deze experimenten opnieuw, en vergeleken dit keer niet de stamgemiddelden met elkaar, maar bekeken de data op individueel niveau door middel van cluster analyse. Hier kwamen twee clusters (groepen) muizen uit naar voren die elk samen groepeerden op elk van de meegenomen gedragsdimensies. De profielen van deze clusters kwamen sterk overeen met de gedragsprofielen van BALB/c en 129 muizen zoals deze werden gevonden in de oorspronkelijke studies: een adaptief en niet-adaptief profiel. Het interessante echter was dat analyse op individueel niveau liet zien dat niet alle 129 muizen een niet-adaptieve reactie vertoonden: een deel van de 129 muizen viel in de groep dieren die wel een adaptieve angstrespons lieten zien. Daarmee liet dit hoofdstuk zien dat binnen 129 muizen subtypen kunnen bestaan die verschillen in hun adaptieve kwaliteit. In dit hoofdstuk kwam verder naar voren dat analyse op basis van de individuele waarden een meer geprononceerde niet-adaptieve angstrespons liet zien dan in eerste instantie werd waargenomen in de oorspronkelijke studies. Hiermee werd gesuggereerd dat het analyseren van data op individueel niveau nieuwe informatie kan opleveren die over het hoofd wordt gezien wanneer men alleen de stamgemiddelden van experimentele groepen vergelijkt.

Het feit dat deze bevindingen waren gebaseerd op een reeds bestaande dataset deed echter de vraag opkomen of deze gevonden variatie daadwerkelijk het gevolg was van inter-individuele variatie in gedragsprofielen, of respons typen, of dat de gevonden variatie wellicht (deels) werd veroorzaakt door het feit dat we onze gegevens baseerden op een serie experimenten die op hun beurt ook weer varieerden in allerlei factoren. Deze data werden bijvoorbeeld verzameld over een tijdsbestek van vier jaar, door meerdere experimentatoren en in verschillende proefdierlaboratoria. Het is bekend dat dit soort factoren ook weer van invloed kunnen zijn op inter-individuele variatie. In **Hoofdstuk 3** werd daarom inter-individuele variatie in adaptief versus niet-adaptief angstgedrag in het mHB onderzocht in dezelfde muizenstammen, maar dan nu in een gecontroleerde dierexperiment.

In deze dierproef werden door twee experimentatoren BALB/c en 129 muizen herhaaldelijk blootgesteld aan het mHB, waarna de individuele responsen op de verschillende gedragsdimensies werden geanalyseerd met behulp van dezelfde cluster analyse. In deze studie werd echter nog een derde stam toegevoegd: C57BL/6, welke bekend staat als niet-angstig. Verder werden naast gedrag ook stresshormonen (bloedplasma corticosteronwaarden) gemeten om te bepalen of eventuele inter-individuele variatie in gedrag zich ook zou vertalen naar soortgelijke verschillen op fysiologisch niveau. Deze stresshormonen werden op drie momenten gemeten om ook hier het verloop van de deze respons te onderzoeken (een week voor de gedragstest, direct na de test, en een week na de gedragstest). De resultaten van **Hoofdstuk 3** bevestigden de bevindingen van het vorige hoofdstuk in die zin dat er op gedragsniveau wederom twee responstypen werden gevonden: dieren die adaptieve angst lieten zien versus muizen die werden gekarakteriseerd als niet-adaptief angstig. De twee eerder gevonden gedragsprofielen werden hiermee empirisch bevestigd voor 129 muizen. Uit dit experiment bleek echter ook dat de twee profielen werden vertoond door individuen van de BALB/c en C57BL/6 stammen. Deze variatie binnen stammen was alleen terug te vinden in gedrag, en kon niet gekoppeld worden aan een verschil in stresshormonen.

Hoofdstukken 2 en 3 brachten op deze manier de inter-individuele variatie in angst-gerelateerd gedrag, maar ook in exploratief gedrag en activiteitsniveau, in kaart bij drie muis-inteeltstammen. In **Hoofdstuk 4** werd deze kennis toegepast om te onderzoeken in hoeverre het meenemen van inter-individuele variatie in gedrag bij het ontwerp van een dierproef effect zou hebben op de uitkomsten van een (gedrags)farmacologisch experiment. Deze studie werd uitgevoerd met in het achterhoofd de gulden regels van een goede proefopzet. Een basisregel van een goed ontwerp van een dierproef is dat alle variabelen van te voren dienen te worden benoemd en tevens onder controle te worden gehouden, behalve die vanwege de behandeling. Een ander fundamenteel principe is dat test en controle groepen bij aanvang van de dierproef zoveel mogelijk op elkaar dienen te lijken, met minimale variabiliteit binnen elke proefgroep. In een eerste fase van dit onderzoek werd gebruik gemaakt van dezelfde proefopzet als in Hoofdstuk 3: BALB/c, 129S2 en C57BL/6 muizen werden op individueel niveau gekarakteriseerd door twee experimentatoren, door ze herhaaldelijk te testen in het mHB. In deze fase werden de gedragsobservaties aangevuld met de bepaling van stresshormonen (op dezelfde wijze als in Hoofdstuk 3). Uit deze karakterisatie kwamen wederom twee responstypen naar voren met een vergelijkbaar profiel als in de eerdere hoofdstukken: een adaptief versus niet-

adaptief gedragsprofiel. Net als in Hoofdstuk 3 kwamen deze subtypen voor in alle drie de muizenstammen, en konden ze niet gekoppeld worden aan variatie in de stresshormonen.

Voor de tweede stap in deze studie werd de karakterisatie van de dieren toegespitst op de drie gedragsdimensies die het grootste deel van de variatie tussen responstypen verklaarden: vermijdingsgedrag, exploratie en locomotie. Fase 2 bestond vervolgens uit een farmacologisch experiment waarbij de effectiviteit van dexmedetomedine als anxiolyticum werd onderzocht op dezelfde muizen als Fase 1. Het experiment was zo opgezet dat de ene helft van elk duo muizen het anxiolyticum dexmedetomedine kreeg toegediend, terwijl de andere helft van dat duo een controle behandeling (fysiologisch zout) kreeg. Elk dier werd daarna geobserveerd in een enkele mHB-trial. Deze duo's werden door middel van een volledig gerandomiseerd blokontwerp gevormd. Dit gebeurde binnen de factoren stam en experimentator, zodat er voor elke stam die werd getest door elke experimentator evenveel duo's waren. Deze duo's werden getest op vier testdagen, waarvan elk testdag gold als een blok. Bij de helft van de duo's binnen stam en binnen experimentator werd vervolgens rekening gehouden met inter-individuele variatie door elk duo te matchen op basis van gewicht én responstype. Bij de andere helft duo's (binnen stam, binnen experimentator) werden de muizen binnen een duo alleen gematcht op gewicht.

Het systematisch meenemen van inter-individuele variatie in responstype in de samenstelling van de behandelingsgroepen leidde inderdaad tot andere resultaten dan wanneer deze variatie niet werd meegenomen. Zo had deze variatie onder andere een uitvergrotend effect op de werking van dexmedetomedine bij gedrag dat was gekoppeld aan activiteit. Tegelijkertijd had deze variatie een dempend effect op de storende invloed van experimentator bij het meten van vermijdingsgedrag.

Dit verschil werd toegeschreven aan het al dan niet meenemen van inter-individuele variatie, aangezien de twee groepen muizen zoveel mogelijk vergelijkbaar waren gehouden op andere factoren. Hoewel de bevinding dat inter-individuele variatie een storend effect heeft op de uitleesvariabelen niet nieuw is, vormt dit experiment naar wij weten wel het eerste onderzoek waarin dit effect empirisch werd aangetoond. Met de bovengenoemde gulden principes van proefopzet in het achterhoofd werd verder geconcludeerd dat het meenemen van inter-individuele variatie in de samenstelling van de behandelingsgroepen de kwaliteit van een dierproef ten goede komt. Immers, test-controle duo's werden nog meer vergelijkbaar door het meenemen van inter-individuele variatie.

Box II beschrijft de verschillen in gedrag en stresshormonen tussen stammen uit Fase 1 van Hoofdstuk 4. Deze gegevens zijn niet meegenomen voor verdere analyse in het artikel dat voortvloeide uit het onderzoek van Hoofdstuk 4 en staan hier ter inzage.

Hoofdstukken 2, 3 en 4 brachten inter-individuele variatie in angst-gerelateerd gedrag bij muizen in kaart. Zoals eerder beschreven is de uiting van angst bij muizen een multidimensioneel fenomeen. Dit multidimensionele karakter is echter niet alleen op gedragsniveau te zien, maar ook bij overerving speelt iets soortgelijks. Onderzoek bij zowel mensen als muismodellen toonde aan dat angst een complexe overerving heeft waarbij meerdere genen betrokken zijn. Deze genen kunnen met elkaar interacteren, maar tevens door allerlei epi-genetische en omgevingsfactoren beïnvloed worden. Deze complexe werking is vervolgens weer van invloed op inter-individuele variatie in de uiting van en ontvankelijkheid voor (pathologische) angst. De identificatie van de genen die betrokken zijn bij deze processen vormt een van de hoofddoelen binnen preklinisch neuropsychologisch onderzoek. Dit soort identificaties begint vaak met het in kaart brengen van zogeheten QTLs; deze afkorting staat voor 'Quantitative Trait Loci'. Een QTL is een feite een gebiedje op een chromosoom waarin zeer waarschijnlijk en gen met een zekere invloed op een bepaald kenmerk ('trait'), zoals angst, ligt. Eerder onderzoek binnen het departement Dier in Wetenschap en Maatschappij (DWM) liet door gebruik van QTL-analyses zien dat de uiting van ongeconditioneerd angstgedrag in het mHB te koppelen is aan specifieke loci, op bepaalde chromosomen. Met name loci op chromosomen 1, 10, 15 en 19 kwamen hier naar voren. In deze studies werd echter niet naar het verloop van een angstrespons gekeken, maar alleen naar de reactie van muizen na een enkelvoudige blootstelling aan het mHB. Met betrekking tot de gevonden inter-individuele variatie uit hoofdstukken 2, 3 en 4 waren wij benieuwd of het verloop van de angstrespons in deze eerder gedane onderzoeken ook te koppelen zou zijn aan deze (of andere) loci en chromosomen. Om dit te onderzoeken analyseerden we in **Hoofdstuk 5** de gegevens uit het eerdere onderzoek opnieuw. De gedragsdata uit dit onderzoek bestond uit een enkele mHB-trial van 5 minuten. Om het verloop over tijd van deze trial te analyseren werden deze data omgezet naar trajectories met vijf tijdspunten, één per minuut. Uit de statistisch-genetische analyse die

werd uitgevoerd in dit hoofdstuk bleek dat chromosomen 10 en 19 konden worden geassocieerd met het verloop van de angstrespons over tijd. Deze chromosomen lijken dus niet alleen betrokken te zijn bij de expressie van angst an sich, maar ook bij het moduleren van het verloop van deze reactie over tijd.

Samenvattend geeft het onderzoek beschreven in dit proefschrift enerzijds een gedetailleerd beeld van de mate en aard van inter-individuele variatie in het verloop van angst-gerelateerd gedrag over tijd binnen drie muis-inteeltstammen. Daarmee biedt het een eerste aanzet tot een beter begrip van de mogelijk verschillende biologische processen die ten grondslag liggen aan inter-individuele variatie in (pathologische) angst bij muismodellen. Dit onderzoek toont echter bovenal de potentieel toegevoegde waarde van het meenemen van inter-individuele variatie in de analyse van onderzoeksresultaten, en het ontwerp van een dierproef. Naast het geven van voorbeelden van hoe deze variatie mee kan worden genomen in de statistische analyse en proefopzet, laat dit onderzoek zien dat inzoomen op deze variatie de kwaliteit van dierproeven ten goede kan komen. Daarmee voegt dit onderzoek een extra dimensie toe aan een wijdere stroming welke nieuwe methoden en perspectieven verkent teneinde de kwaliteit en reproduceerbaarheid van dierexperimenteel onderzoek te vergroten.

Appendices

Dankwoord

Ik zou zo nog een boekje kunnen schrijven over dit alles. Sterker, ik héb nog een boekje vol geschreven en dat is allemaal terecht gekomen in een groot bestand genaamd ‘Manuscript_dump’. Het mag allemaal geen naam meer hebben. Dit is het geworden, en ik ben ontzettend trots op het resultaat. Niet in de laatste plaats omdat dit traject bijzondere wendingen heeft gekend waarbij verschillende ideeën, richtingen en mensen de revue zijn gepasseerd. Meer dan het onderzoek-naar-individuele-verschillen-bij-muizen is dit project een zoektocht geweest naar hoe je die verschillen moet kwantificeren, en hoe je deze kunt inzetten, maar ook een oefening in het sturend houden van een project op momenten waarop de sturing noodgedwongen even afwezig was. Of waarbij andere gebeurtenissen in het leven je raken. Los van het wetenschappelijke gedeelte heb ik veel geleerd, over mezelf, over verandering, maar ook over vertrouwen, veerkracht en houvast. En hoe je altijd vertrouwen moet houden op een goede afloop. Dit boekje is daar een mooi voorbeeld van.

Ik had dit onderzoek nooit kunnen doen zonder de geweldige supervisie en support van een flink aantal mensen. Allereerst **Saskia**, eerst betrokken als co-promotor, en de laatste jaren als mijn promotor! Dankjewel voor alle input, begeleiding en aanmoediging tijdens dit traject. Ik wist mij altijd gesteund en gewaardeerd door jouw begeleiding. Je gaf me de vrijheid mijn nieuwsgierigheid achterna te lopen, en wist tegelijkertijd te sturen op een geslaagde afronding. Hoe druk je het soms ook had, je maakte graag tijd voor me en wist het onderzoek altijd weer een stapje verder te tillen met je opbouwende feedback.

Datzelfde geldt voor jou **Hein**, sterker – ik zou niet weten waar ik was geweest zonder jouw supervisie als co-promotor. Wat hebben we veel meegemaakt in zeven jaar! Toen ik voor het eerst bij je langskwam in aanloop naar dit project vroeg je hoe het zat met mijn kennis van statistiek. Ik antwoordde iets in de trant van mja.. de basis moet zeker lukken. Little did I know. Little did we both know denk ik, want ik vermoed dat dit project voor jou soms ook onbekend terrein is geweest van tijd tot tijd. Met het wegvallen van Frauke, de initiator van dit project, was het soms zoeken naar de beste manier om recht te doen aan alle plannen die er lagen, en tegelijkertijd er een eigen draai aan te geven. Die draai komt voor een groot deel uit jouw koker. Wat heb ik veel van je geleerd. Niet alleen op het gebied van statistiek en experimenteel design. Je gaf ook het goede voorbeeld op het gebied van good scientific practice en daagde me uit om mijn keuzes voor een bepaalde strategie te kunnen toelichten en verantwoorden. Je deur stond altijd open, ook online. Ik denk met veel plezier terug aan alle keren dat ik ‘even’ langs kwam met een vraag en we vervolgens urenlang aan het sparren waren over het opzetten en ontwerpen van de experimenten, over statistiek of over waar ik dan ook tegenaan liep. Niet zelden ging het dan ook over allerlei andere dingen. Dankjewel voor al die onvoorwaardelijke support.

Mijn dank gaat ook uit naar **Hans**, voor het tijdelijk inspringen als interim-promotor. Dankzij jouw betrokkenheid heeft dit project op het juiste moment een zet in de goede richting gekregen waardoor we weer verder konden.

Natuurlijk hoort ook **Frauke**, de initiator en oorspronkelijke promotor van dit project, in dit rijtje. Wat ben ik blij dat je me destijds de kans hebt gegeven om dit project op te pakken. Het was een long-shot, ik kwam bij je langs om eens te kletsen en ineens hadden we tot de verbazing van (denk ik) ons allebei een plan aan het eind van het gesprek. Dit project kwam uit jouw koker, en jij zag er de potentie van in. Jouw enthousiasme en vertrouwen maakten dat ik – ook op de momenten dat ik me afvroeg waar ik in hemelsnaam mee bezig was (en die waren er best) – altijd zelf nog in kon blijven geloven. Het is een belangrijke motor geweest om dit project af te maken. Wat zou ik je graag dit eindresultaat laten zien. We hebben ons best gedaan recht te doen aan je plannen, aan wat je voor je zag met dit project.

Janneke en **Maaike**, mijn paranimfen! Voor jullie allebei geldt dat ik me vaak WWMD of WWJD (What Would Maaike/Janneke Do?) afvroeg als ik ergens tegenaan liep. Of het nou om iets inhoudelijks ging, andere PhD-gerelateerde zaken, de beste koffie van de campus of gewone dingen des levens – bij jullie kon ik altijd terecht om even mee te denken en te sparren, met de juiste dosis humor. Het leverde me altijd iets op waarna ik weer met frisse moed (en de juiste koffie) verder kon. Buiten het werk hadden we elkaar ook meer dan genoeg te vertellen, en ik ben dankbaar dat we dat zijn blijven doen. Lieve **Janneke**, mijn partner-in-crime op gezamelijke summer schools en conferenties, maar ook gewoon als (wannabe, ik) hondentrainers op een donker veldje in Den Haag, biertjes en etentjes na het werk en dan rennend door Utrecht CS naar de trein. Als bonus waren we de laatste jaren zelfs kamergenoten bij DWM. Ik bewonder je vrolijke, aanstekelijke enthousiasme, je betrokkenheid bij de mensen om je heen en je kennis en gevoel voor diergedrag. En je talent voor timing: De laatste twee jaar, allebei vanuit huis, checkte je regelmatig even hoe het ging

– vaak net op de momenten waarop ik dat wel even kon gebruiken. Je was daarmee een belangrijke lifeline tijdens het thuiswerken. Lieve **Maaike**, als 'de andere PhD van Hein' lag jouw project nog het meest dichtbij dat van mij. Ik heb dan ook heel erg veel aan je gehad qua advies, tips and tricks en uitleg. Jouw proefschrift is een voorbeeld geweest tijdens het schrijven van dit boekje. Ik bewonder je talent om bij jezelf te blijven en je niet gek te laten maken. En je kennis en inzicht over de meest uiteenlopende dingen, van wetenschap en actualiteit tot showbizz en achterklap ☺. Niet zelden liepen de tranen over onze wangen van het lachen. Ook met jou heb ik meer dan veel herinneringen aan al die keren dat de werkdag overging in een borrel of burgers bij de Basket. Dank jullie allebei voor zoveel gezelligheid, betrokkenheid en plezier. Ik ben vereerd en blij dat jullie naast mij willen staan bij mijn verdediging.

Jan, met je brede kennis van statistiek vond je voor ons een methode die precies aansloot bij onze wensen. Hierdoor konden we verder met onze vragen. Ik ben er trots op dat je co-auteur van mijn eerste paper wilde zijn.

Amber, **Hetty**, **Maaike** en **Marijke** Laarakker, de gegevens van jullie studies vormen een belangrijke component van dit boek, dankjewel dat ik deze mocht gebruiken!

Melissa, Suzanne en **Marieke**, zonder jullie geweldige inzet bij het experimentele werk zou dit proefschrift niet zijn geworden wat het nu is. Dankjewel, het was een plezier jullie te mogen begeleiden. Suzanne, het kon natuurlijk niet anders dan dat ik de geweldige titel van je master thesis graag wilde lenen voor dit proefschrift, dankjewel voor je toestemming hiervoor!

AM, **José**, **Susanne**, **Judith**, **Lisa** en **Eline**, dankjewel voor al jullie hulp bij het draaien van de experimenten. Tegen de tijd dat ik begon met de studies liep ik al even rond bij DWM waardoor het voor iedereen (inclusief mijzelf) even duurde voordat duidelijk werd dat ik eigenlijk nog weinig lab-ervaring had. Met jullie geduld, training en support is dat helemaal goed gekomen! Ik heb veel van jullie geleerd. Datzelfde geldt voor **Marijke**, als kamergenoot van de M-kamer was je vaak vraagbaak nummer één voor alles wat te maken had met het opzetten van de dierstudies. Altijd nam je geduldig de tijd om me alles bij te brengen over werkprotocollen, wie wat waar op het GDL en veel meer.

Nicky, **Anja** en **Wout** van het GDL, dank voor al jullie hulp bij de experimenten en de goede zorg voor de dieren. **Monica** voor de check-ins van tijd tot tijd in de laatste thuiswerkperiode, en **Mechteld** voor een last-minute printopdracht waar je u tegen zegt. En **Guus,** dank voor het ontwerp van de geweldige cover van de boekje en de mooie lay-out, ik ben heel blij met het resultaat.

Dat ik veel plezier heb gehad bij het onderzoek van dit proefschrift komt voor een niet onbelangrijk deel door de vele (ex-)collega's bij DWM. Zeven jaar is best een tijd, en ook al vond de laatste tijd van deze periode thuis achter de computer plaats, ik heb veel met jullie meegemaakt en ontzettend veel gelachen. **Hetty**, **Marijke**, **Susanne**, **Judith**, **Eline**, **Lisa**, **Saskia** Kliphuis, **Maëva**, **Elske**, **Janneke** Arts, **Vivian, Heidi:** van gezamelijke tripjes tot escape rooms, borrels en bbqs: dankjewel voor al die gezelligheid en jullie betrokkenheid. Ook de overige (ex-) collega's van DWM, **Monica**, **Mechteld**, **Irene**, **Pim**, **Jan**, **Jan** Langermans, **Louk**, **Tara**, **Ate**, **Paulien**, **Petra**, **Bas**, **Mona**, **Theo**, **Nienke, Claudia**, **Marjan**, **Joanne**, **Chantal**, **Matthijs**, **Maite**, **Nelleke, Franck**, **Joachim**, **Ellen**: dank voor de prettige werksfeer, mede daardoor was het een plezier om naar mijn werk te gaan.

Een aantal collega's wil ik nog even in het bijzonder noemen. Allereerst natuurlijk de roomies van de 'M'-kamer: **Maaike**, **Marsha**, **Marcia**, **Marijke**, en later **Malou** en **Emmy**. Ik weet niet in hoeverre de rest van de gang altijd gelukkig was met het geklets dat soms uit onze kamer kwam maar wat was dat een leuke tijd. Hoe fijn om het lief en leed van een PhD-traject te kunnen delen, maar ook alle andere hoogte- en dieptepunten uit het leven daarbuiten! **AM**, je bent van alle markten thuis en onmisbaar: hechtinstructies, een stoomcursus tapdansen (kunnen wij!), maar vooral stond je altijd klaar voor fijne gesprekken, goeie grappen of de juiste woorden op het juiste moment. En dan **Marsha**, ook jij bent onlosmakelijk met de 'M'-kamer verbonden, vanaf dag één heb je me een ontzettend welkom gevoel gegeven en ook jij stond altijd klaar voor een luistered oor, een GT en veel gezelligheid buiten het werk. **Esther**, de manuscript crunchende PhD (lang geleden) aan de andere kant van het raam, ik herinner me de gezelligheid in Cairns en in Utrecht. **José**, een berichtje was nooit ver weg, ook jij bedankt voor de betrokkenheid en de vele gezellige momenten.

Janneke, **Maryse** en **Janna**, jullie werden mijn tweede thuis bij DWM. **Maryse**, je hebt me zoveel tips en tricks ingefluisterd en bent een belangrijke bron geweest in de motivatie om dit boekje af te ronden, met je aanmoedigende

berichtjes van tijd tot tijd. Dankjewel daarvoor! Lieve **Janna**, dit boekje is volledig geschreven met behulp van jouw onvolprezen planningoverzicht. Als roomie genoot ik van je fijne humor, je scherpe blik en je gezelschap, die herinneringen zijn me zeer dierbaar.

Naast alle betrokkenen op werkgebied wil ik ook graag mijn vrienden en familie bedanken.

De vrienden uit Rotterdam, mijn oudste vrienden maar still going strong. De clubjes uit Nijmegen, en dan natuurlijk ook nog de lieve mensen uit Den Haag en Rijswijk! Veel van jullie zijn over de hele wereld uitgewaaierd. En dan ben ik ook nog eens van de ene kant naar de andere kant van het land verhuisd. We zien elkaar niet meer zoveel als vroeger, maar voor ieder van jullie geldt dat we als vanzelf de draad weer oppakken wanneer we elkaar zien en dat betekent veel voor me. Jullie hebben ieder waarschijnlijk meer dan je je realiseert bijgedragen aan dit boekje in de vorm van ontspanning, welkome afleiding, een luistered oor, motivational speech of gewoon gezelligheid. Ik ben dankbaar voor jullie vriendschap.

Mijn schoonouders **Bob** en **Lidwieke**. Jullie zijn door de jaren heen altijd betrokken geweest bij het wel en wee van mijn PhD. Net zoals jullie altijd klaar staan om ergens in te springen, of gewoon om even aan te haken. Lidwieke, een speciale dank gaat naar jou, voor al die keren dat we een beroep op je konden doen om op de kinderen te passen. Ik durf wel te zeggen dat dat een belangrijke bijdrage heeft geleverd aan de totstandkoming van dit boekje, zeker de laatste twee jaar. Ik ben dankbaar zulke lieve en fijne schoonouders te hebben. **Wouter**, **Suzanne** en **Ked**, dat geldt natuurlijk ook voor jullie, dankjewel voor jullie interesse, berichtjes en gezelligheid door de jaren heen!

Suzie, mijn liefste zus. Wat hebben we een boel meegemaakt in de afgelopen jaren, en dan moest dit boek ook nog af. Je kent me door en door en weet me als geen ander te motiveren, aan het lachen te maken of te bellen over niks. Dankjewel voor al die keren dat je me aanhoorde, wist op te beuren, of meedacht. Ik zou niet weten waar ik zonder je zou zijn. Dit boekje ook niet. **Thijs**, wat ben ik blij met jou als lieve schoonbroer. Dankjewel ook voor jouw berichtjes, telefoontjes en betrokkenheid.

Lieve **pap** en **mam,** mijn basis, mijn thuis. In dat thuis is de laatste jaren veel veranderd, maar de basis blijft. Jullie hebben mij altijd gestimuleerd op mijn gevoel te durven vertrouwen en het beste uit mezelf te halen. Ik ben dankbaar voor jullie onvoorwaardelijke liefde, steun en vertrouwen. Ook jullie stonden altijd klaar voor een luistered oor, of een oppas hier of daar. **Pap**, een van de eerste dingen die je zei was hoe jammer je het vond dat je dit moment niet meer mee ging maken. Ik ben dankbaar voor de tijd die er nog wel was, en waarin we alles tegen elkaar hebben kunnen zeggen. **Mama**, je bent de liefste oma voor de kinderen en een voorbeeld voor mij als mama. Je aanwezigheid op de dag van de promotie is me dubbel zo lief. Waar ik ook ben – jullie zijn altijd bij me.

De totstandkoming van dit boek heeft – zeker de laatste sprint in covid tijd – de meeste impact gehad op ons gezin. Mijn grootste dank gaat dan ook uit naar jullie, lieve **Jasper**, **Jetje** en **Sjaak**. Lieve **Jet**, jouw vrolijkheid en liefheid zorgden voor de nodige relativering en afleiding van het schrijven. En je grapjes verdienen een eigen proefschrift. Hoe heerlijk dat het nu klaar is, dat ik zonder bezwaard geweten kan roepen dat ik graag met je ga kleuren! En lieve **Sjaak**, jouw komst deze zomer maakte ons gezin compleet. Nu al genieten we van je heerlijke lach en je lieve blik. Ik kan niet wachten verder te ontdekken wie je bent en te zien wat je voor ons in petto hebt. Jullie brengen zoveel vrolijkheid, plezier en liefde in ons leven. Ik hou van jullie, het is mijn grootste geluk jullie mama te zijn.

En tot slot slimme, dappere **Jasper**, mijn liefste lief. Wat zit er veel van jouw geduld, rust, positiviteit, relativering, en altijd aanwezige bereidheid om mee te denken in dit proefschrift. Je moet er af en toe horendol van zijn geworden, maar je gaf me nooit het gevoel dat het teveel was. Net zoals je dat eigenlijk met alles doet, en dat vind ik heel bijzonder. We hebben veel samen meegemaakt de afgelopen jaren. Zowel de hoogte- als de dieptepunten doe je net zo hard mee, het maakt me dankbaar dat we dat samen kunnen beleven en ondanks alles altijd wel iets te lachen hebben. Ik kijk uit naar de toekomst met ons vieren, én naar meer tijd voor ons tweeën. Ik hou van je, ook dat is mijn grootste geluk – jouw meisje te zijn.

Curriculum Vitae

Marloes Hieke van der Goot was born in Arnhem on May 27th 1979. After obtaining her athenaeum degree at Melanchthon Scholengemeenschap in Rotterdam she moved to Nijmegen to study Psychology at the Radboud University. Following a specialization in Biological Psychology she conducted her masters' internship at this department, with a project on response variability through operant learning in human subjects under the supervision of dr. Roald Maes. The specialization in Biological Psychology sparked her interest in the study of animal behavior, which resulted in a literature master thesis on the function of social play fighting in rats under the supervision of dr. Maes.

Her academic background was further shaped and broadened through her activities as a research assistant at the Max Planck Institute (MPI) for Psycholinguistics in Nijmegen. She worked on various projects at the intersection of psycholinguistics, comparative cognitive psychology and cognitive neuroscience. She studied the evolutionary origins of human social behavior and communication through comparative research in great apes and human infants. In collaboration with the Max Planck Institute for Evolutionary Anthropology in Leipzig she conducted an independent research project on non-verbal communication in great apes and pre-linguistic infants, under the supervision of prof. dr. Michael Tomasello and Dr. Ulf Liszkowski.

In 2014 she started her PhD project under the supervision of dr. Hein van Lith and prof. dr. Frauke Ohl†, and later by prof. dr. Saskia Arndt, at the Department of Animals in Science and Society at the Faculty of Veterinary Medicine in Utrecht. She investigated inter-individual variability in anxiety-related behavior in inbred mice and its implication for the interpretation of research outcomes, the results of which are written up in this dissertation. In addition, Marloes was active as a faculty representative of the PhD-council of the Graduate School for Life Sciences Program Clinical and Experimental Neuroscience. She will continue her career in the direction of bio-informatics.

List of publications

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

Van der Goot, M.H., Keijsper, M., Baars, A., Drost, L., Hendriks. J., Kirchhoff, S., Lozeman-van 't Klooster, J.G., van Lith, H.A., & Arndt, S.S., 2021. Inter-individual variability in habituation of anxiety-related responses within three mouse inbred strains. Phys. Behav. 239, 113503, https://doi.org/10.1016/j.physbeh.2021.113503.

Van der Goot, M.H., Boleij, H., van den Broek, J., Salomons, A.R., Arndt, S.S. & van Lith, H.A., 2020. An individual based, multi-dimensional approach to identify habituation patterns in inbred mice. J. Neurosci Methods, 343, 10880, https://doi.org/10.1016/j.jneumeth.2020.108810.

Van der Goot, M.H., Tomasello, M., & Liszkowski, U., 2013. Differences in the Nonverbal Requests of Great Apes and Human Infants. Child Dev. 85 (2), 444-455, https://doi.org/10.1111/cdev.12141.

Maes, J.H.R. & **Van der Goot, M.H**., 2006. Human operant learning under concurrent reinforcement of response variability. Learn. Motiv. 37, 79 – 92, https://doi.org/10.1016/J.LMOT.2005.03.003.

You can have the most beautifully designed experiment with the most carefully controlled variables, and the animal will do what it damn well pleases.

Harvard Rule of Animal Experimentation