Published in final edited form as:

Curr Protoc Mouse Biol. ; 6(4): 364–379. doi:10.1002/cpmo.17.

Methodological Considerations for Optimizing and Validating Behavioral Assays

Stacey J. Sukoff Rizzo[1] and **Jill L. Silverman**[2]

[1]Mouse Neurobehavioral Phenotyping Facility, Center for Biometric Analysis, The Jackson Laboratory, Bar Harbor, Maine

[2]MIND Institute and Department of Psychiatry and Behavioral Sciences, School of Medicine, University of California Davis, Sacramento, California

Abstract

Preclinical animal models are indispensable tools for translational research for which behavioral characterization and phenotyping are essential to testing hypotheses and for evaluating the potential of novel therapeutic agents to treat diseases. The methods employed for comprehensive behavioral phenotyping and pharmacological experiments are complex and should be conducted exclusively by trained technicians with demonstrated proficiency. The ultimate goal is to identify disease-relevant and translational behavioral endpoints that are robust, reliable, and reproducible, and that can be employed to evaluate potential of novel therapeutic agents to treat disease. The intent of the present article is to provide a pragmatic outline for establishing and optimizing behavioral assays and phenotyping batteries, ensuring that the assays and the data are reliable such that they can be reproduced within and across technicians and laboratories and, more importantly, that the data is translatable to the clinic.

Keywords

assay validation; animal models; behavior; behavioral testing; mouse; pharmacology; phenotyping; reproducibility

INTRODUCTION

The Behavioral Testing Environment

Designing and converting a laboratory space for behavioral testing is not trivial; it is crucial to identify adequate space to place the behavioral testing equipment, understand the limits of the testing environment, and in particular ensure confidence in the sensitivity of the assay to detect the expected endpoints, particularly when the model system has inherent variability of a live, behaving animal. It is not as simple as purchasing the behavioral equipment, placing the equipment in any available laboratory space, placing the mouse in the equipment, and pressing “start” on the computer to record and analyze the data. Rather, it is essential to understand the limitations of the testing environment where the equipment is located and whether the environment is sufficiently optimized to the extent that it is sensitive to detect the expected outcomes. In addition, it is crucial that the technician is proficient in running the test itself. In general when identifying appropriate space intended for sensitive behavioral testing, it is best to avoid having procedure space in high traffic areas or in areas in proximity to cage wash facilities, elevator shafts, or restroom facilities, in order to minimize random disruptions of noise and vibration. While there are fairly simple ways to minimize noise, minimizing of vibration is an extremely important consideration as it is well documented that high vibration levels can impact breeding and pup survival (Rasmussen et al., 2009). A consistent and rigorously environmentally controlled procedure space is a major factor in achieving reliable, reproducible results. Several excellent publications have provided guidance on optimizing specific behavioral testing protocols (Crawley and Paylor, 1997; Crawley, 2007; Buccafusco, 2009; Wahlsten, 2010; Fuchs et al., 2011).

Conceptualizing Assay Validation in Behavioral Testing

Much has been written on the standardization of behavioral testing methods (Crabbe et al., 1999; Würbel, 2000; van der Staay and Steckler, 2002; Würbel, 2002; Wahlsten et al., 2003; Schneider et al., 2006; Mandillo et al., 2008), and while there are inherent differences in laboratory space (variations in testing equipment, housing conditions, and skillfulness and proficiency level of technical staff, among others), the great equalizer across these many often uncontrollable and/or unknown variables is the ability to demonstrate, under the given laboratory conditions, the test is indeed sensitive enough to detect the expected behavioral changes (i.e., assay validation). In this respect, far before any experimental unknowns are tested, initial experiments should be conducted to ensure the ability, reliability, and sensitivity of the assay being established to produce the expected baseline results when a positive or known standard is evaluated. Irrespective of how much standardization is even possible, the only way to ensure that the test is op-timized for detecting the expected behavioral changes and to confirm the proficiency of the technician is to demonstrate that under the conditions being tested that a positive control can produce the expected result. For example, if the aim is to set up a test sensitive for detecting an anxiolytic-like effect of a novel compound, then the technician should be able to demonstrate the ability of a standard anxiolytic agent (e.g., diazepam) to produce an anxiolytic-like effect. In the absence of this, it will be a challenge to understand whether the experimental variable (the test compound) fails to produce an effect in the assay or whether there was fault with the testing environment or the technician’s ability to conduct the test properly, which includes a number of variables ranging the gamut from handling, restraint, and injection skills to careful data analysis. This, however, should not be confused with the concept that a novel mechanism of action may not produce behavioral effects identical to the effects of known standards from which the behavioral assay may have been optimized for, but rather provides the confidence that the test was conducted under the optimal conditions for which it was established (Tricklebank and Garner, 2012).

The Pillars of Reproducibility

A well-conceived experimental design should aim to be reproducible, taking into consideration the application of several key principles that minimize as many environmental variables as possible as well as eliminate any potential bias (Unger, 2008; Kilkenny et al., 2009; Kilkenny et al., 2010; Landis et al., 2012; Oswald and Balice-Gordon, 2014). These pillars of reproducibility include blinding, randomization, counterbalancing, suitable sample sizes, and the inclusion of appropriate controls (Fig. 1). The inherent variability of live, behaving animals is the greatest factor for considering an experimental design which means it is crucial that every other factor (e.g., environmental, experimental) should be controlled as rigorously as possible so that only the unknown or experimental reagent becomes the single variable being evaluated.

Blinding: In a blinded experiment, at minimum the technician responsible for directly evaluating the behavior and analyzing the data should not be aware of the treatment groups. In some cases this may be challenging if there are visual clues (e.g., coat color, solution color) that may render the study infeasible for blinding. In these cases the analysis and interpretation of the data should be performed by an independent technician who is only revealed the treatment code after the data have been interpreted. The methods of which blinding was achieved should be clearly reported with substantial details in the experimental methods text.

Randomization and counterbalancing: Test subjects should be randomly assigned to treatment groups. When baseline testing is conducted such that drug treatment will be compared relative to pretreatment baselines, then considerations should be made for counterbalancing (i.e., performance levels, body weights) evenly across treatment groups so as to not bias high or low performers into a single group. Thus, each treatment group should be evenly represented by low and high performance levels such that there should be no statistical differences across groups prior to initiating drug treatment. The principles of randomization and counterbalancing should also be applied across testing sessions, time of test day, and multiples of testing equipment, as well as considerations for assigning treatments within a group housed cage. For example, if testing requires several days to complete due to large group size, limited apparatuses, equipment, and software instrumentation, then representative subjects from each treatment group should each be tested across the testing days. Not only does this minimize bias in the experiment, but it also proactively plans against losing an entire treatment group in the case of unplanned events that would require termina-tion of an experiment (e.g., power outage, fire alarm).

Controls: For compound screening, at minimum a vehicle control should be included within the experimental design. Irrespective of the consistency of historical control data in the laboratory, it is important to understand relative change within an experiment, whether that is relative to a vehicle-treated control or relative to a wild-type control in a phenotyping experiment. With respect to preclinical pharmacological assays, the vehicle control group should receive the same vehicle that the test compound was formulated in and matched for excipients and pH levels. Handling-induced and injection-related stress is an important variable that can contribute to behavioral outcomes. Identical treatment of controls—with the exception of the test compound—will provide confidence in the interpretation of the result that it was indeed due to the test compound and not due to any other contributing or confounding variable.

Sample size: Group sizes of 10 to 20 per sex, per genotype/treatment are typically the minimal sample sizes required to achieve statistical significance in a given assay based on behavioral experience and previous power analyses. Effects of genotype and sex should be evaluated using multi-factor analysis of variance (ANOVA). Significant ANOVAs should be followed with Tukey’s high significant difference test or other appropriate post hoc tests to identify specific differences between groups (Silverman and Crawley, 2014). It is not appropriate to generate data with small sample sizes (e.g., $n = 2$ to 8 per sex/per genotype or treatment) in independent experiments and then combine them with other small samples from separate experiments to increase power. Rather, initial findings from small cohorts can be considered pilot data and used to generate power calculations for follow-up experiments, with a second set of appropriately powered experiments planned in an independent cohort and executed to confirm results, prior to publication. It is also not appropriate to combine sexes within an experiment unless statistical analyses are performed to demonstrate a lack of an effect of sex, and methods for how the data were analyzed should be transparently reported. If the data cannot be confirmed in one’s own laboratory, then it is less likely that it can be reproduced under different conditions in another laboratory.

While the application of these pillars of reproducibility will provide for a more reliable assessment of the behavioral phenotype and of the behavioral effects of a compound being tested, it should be clear that convergent data from multiple behavioral tests as well as correlating biochemical data are important for strengthening the reliability of the mouse model being evaluated or the compound being tested and its translational utility (Cryan and Mombereau, 2004; Tricklebank and Garner, 2012; Rizzo et al., 2013).

Technical Proficiency

A mastery of conducting sensitive behavioral tests can be met by ensuring that the technician is trained if they can accomplish reproducing published data sets (either test compounds or published phenotypes in well-described and reproduced mouse models which can serve as positive controls). To ensure proficiency, not only should the tech-nicians be able to accomplish this, but they should be blind to the treatment groups or genotypes such as to eliminate any potential bias as well as provide confidence in the technician’s proficiency. This is the ultimate test of a technician’s proficiency; failure to reproduce positive control data when all variables are known should caution the investigator that an unknown (a mutant mouse line or a novel pharmaceutical compound) is not ready to be tested, as there are no ideal standards of comparison. Training of technical staff to a level of proficiency should be budgeted appropriately, both with respect to study costs and time, as several attempts to achieve successful results may be required. The benefit of this is the ability to demonstrate that the assay has been optimized for being able to detect the expected behavior, as well as an increased confidence level of the technician and their colleagues to trust the data and minimization of any questions when an unexpected outcome occurs and the technician’s proficiency is challenged. Examples of commercially available reagents (both positive control compounds and mouse models) that can be used to demonstrate assay validation and confirm proficiency of the technician are provided in Table 1. Further, allotment of this training time allows for the technician to master the high level of attention to detail and multi-tasking required to successfully execute a behavioral study given the exquisite timing of second by second events inclusive of time to set up and habituate the subjects, time to place and remove the mice from the testing arena, time to observe the mice in the arena, time to clean the testing arena between subjects, and time to complete the experiment, clean up and rehouse the mice, and analyze the data.

All Mice Are Not Equal

The selection of mouse strain is a fundamental consideration when optimizing behavioral assays. Regardless of mouse model employed (i.e., genetic mutant, disease model), it is critical to understand how the background strain, and both sexes, perform in the behavioral task prior to moving forward with the experimental cohort of mice (Crawley et al., 1997). While historically the C57BL/6 mouse has been primarily used in behavioral testing given its frequent application as a background strain for genetically engineered models, its generally consistent performance across behavioral assays, and its ease of accessibility, it is important to be aware that inbred strains such as the C57BL/6 strain have substrains (e.g., C57BL/6J and C57BL/6N) that have well-reported behavioral divergences and baseline values across assays (Bryant et al., 2008; Simon et al., 2013). Investigators should exhibit due diligence and report knowledge of the substrains, as well as number of generations bred and backcrossed, particularly when using hybrids. Further, wild-type controls are not inbred mice and should not be expected to perform identically to performance levels of historical data from the inbred strain. Reagents that contribute to the generation of genetically manipulated mice (e.g., neo-cassettes, cre-drivers, flagging techniques), could very well influence behavioral responses of the wild-type controls, especially when number of backcrossings through subsequent generations is minimal. A working knowledge of the background strains or substrains, and their basal behavioral responses in each behavioral test, is critical for understanding whether findings in the experimental cohort are related to the inherent behaviors of the background strains themselves. Therefore, if this is not the case, a baseline cohort of full Ns, should be evaluated to have ideal or expected values for each assay.

Confounding Behavioral Responses

The impact of a competing or confounding behavior on the behavioral endpoint being evaluated should not be underestimated. In certain strains, behavior may be less than ideal or not feasible at all. For example, mice on a C3H or FVB inbred background carry a retinal degeneration allele and have visual impairments with age (Schellinck et al., 2010). Strains with visual impairments may not be useful for cognitive tests that employ visual cues as reference stimuli, and, further, it is well reported that blind mice tend to be hyperactive, a behavior that is often a confound in many behavioral tasks (Dyer and Weldon, 1975). Hyperactive mice also demonstrate increased ability to maintain their balance on the rotarod, a behavioral effect that can be mimicked by administration of a stimulating dose of amphetamine to nonhyperactive mice. Mice that have hearing impairments (either from birth or through aging) may not be useful for tasks that employ audio cues. Mice with impaired olfaction may have reduced social behavior or may not be appropriate for food-motivated tasks. Just as it is important to understand the limitations of a behavioral task itself, it is important to investigate, acknowledge, and report the limitations of the mouse model being tested so as not to be myopic in the interpretation of the data (Table. 2).

The potential confounds of behavioral responses are greatly increased when a drug is being tested. A nonspecific effect of virtually all drugs at an excessive dose level is sedation. Sedative behavior confounds nearly all other behavioral responses since the mouse's physical, active engagement is required in behavioral tasks. It is therefore important that a test compound's pharmacokinetic properties (e.g., half-life, time to maximal concentration) are known such that an appropriate dose range, route of administration, and pretreatment time can be selected. It is also important to be aware of the dose range which is specific to the biological target relative to the nonspecific dose that produces sedation or other adverse effects (Rizzo et al., 2013). Thus it is critically important to be cautious in interpreting data for the endpoint of a behavioral domain that could well be influenced by another competing, dominating behavior. Aside from the test compound potentially producing adverse behavioral effects, many reagents commonly used as vehicles for in vitro or biochemical assays (e.g, DMSO, Tween, ethanol) produce behavioral effects on their own (Castro et al., 1995; Lin et al., 1998; Colucci et al., 2008; Rivers-Auty and Ashton, 2013).

Although behavioral investigators and personnel are aware of the need to control the macroenvironment of the animal room within our laboratories, it is also important to control the microenvironment in animal cages from birth to testing or from vendor to laboratory. The microenvironment within the cage may influence the biological response of the animal test system. Influences that may add variability include housing methods, light, bedding, noise, diet, transportation, temperature, chemicals in feed and bedding, humidity, air quality, ventilation, water treatment, animal handling, vibration, and caging and accessories (Everitt and Foster, 2004). One major trend is to enlarge cages and add objects and complexity for the purpose of environmental enrichment to address animal welfare concerns (Olsson and Dahlborn, 2002). Other animal welfare considerations, such as the use of social housing, have not been rapidly embraced within community due to rising costs, but these changes are now slowly occurring along with validation of methods (Turner et al., 2003). Diet and sterility of vendor and vivaria facilities all influence gut microbiome, which is been increasingly shown to affect behavioral outcomes. Pioneering studies on gut-brain microbiomes in animal models have determined there are plausible impacts of microbiomes on animal behavior (Cryan and Dinan, 2012).

Designing of Comprehensive Testing Batteries

Mouse models are invaluable tools, and every care should be taken to maximize their value. It is not only practical, but common practice for subjects to be tested through comprehensive behavioral testing batteries so as to minimize total number of naïve mice required for individual tests. It is advisable, however, that comprehensive testing batteries should be designed such that mice are evaluated first in the least invasive assays through to more invasive tests, with those tests that have potential to be stress-inducing occurring towards the end of a testing battery (McIlwain et al., 2001). In general, a 1 to 2 day rest period is a sufficient rest interval for mice between tests (Paylor et al., 2006). It has been reported that testing order can influence results of subsequent tests; therefore, subsequent or confirmation cohorts of mice should be treated identically in order to be able to reproduce the initial results (McIlwain et al., 2001). When it is not known how order of testing may influence a subsequent behavior, then it is prudent to investigate this at minimum in the background strain. In some cases it may be of interest to assess basal behavioral responses in the early part of the testing battery and then re-evaluate the same behavior at the conclusion of a testing battery to understand how the frequent handling and exposure to additional testing may impact the behavior relative to the initial basal response. For example, differences in basal anxiety response in relatively naive mice at the beginning of a testing battery may change following repeated handling and testing in other behavioral assays. It is also possible to schedule longitudinal phenotyping in the same cohort of animals, particularly when the disease endpoints are relevant to age of onset. Mouse models of neurodegenerative disorders are often evaluated at a predisease stage (~ to 6 months of age) and then retested at 9 months, 12 months, and later age time points. Importantly, the details of not only the methods for the behavioral assays, but also the order of testing, inter-testing interval, and age of mice should be provided in publications such that the data can be independently reproduced.

All protocols using live animals must first be reviewed and approved by an Institutional Animal Care and Use Committee (IACUC) or must conform to governmental regulations regarding the care and use of laboratory animals.

MATERIALS

Animals

1. The selection of an appropriate strain, age range, and of both sexes should be relevant for the hypothesis being tested. While historically male test subjects have been primarily used for behavioral testing in order to avoid the additional variability associated with the behavioral effects of the female estrous cycle, considerations should be made to include female mice, although a priori sex should be included as an independent variable. Irrespective of the behavioral changes associated with the estrous cycle, male and female mice have divergent behaviors in different behavioral assays and crucially may vary in sensitivity to drug treatment and dosage. When males and females are included in a behavioral test, extra caution should be taken to eliminate female scent from testing equipment when a female precedes a male. Excessive sniffing may result in reduced exploration and reduced performance levels. It is acceptable, given that sex is analyzed as an independent factor, that females are tested only after males have completed testing or on a separate test day to avoid the potential influence of reproductive hormones on the males' behavioral responses.

2. Age at testing is an important consideration and is an essential component of the detailed methods. Typically, behavioral studies that are not including a developmental battery are initiated at adult age (≥8 weeks of age). If the testing requires developmental assessments, than the test battery may be designed to evaluate mice from postnatal day 1 (PND1) through early development into wean age (3 to 4 weeks) and beyond. It should be noted that intervening with maternal care by manipulating the pups during the early pre-wean period, may in itself influence development and may need to be controlled systematically by a sham group of pups that are not handled through this development period. In general

for pharmacology studies, unless the clinical plan specifically targets dosing in a pediatric population, then dosing studies should be limited to adult mice (≥ 8 weeks of age) to ensure that the drug itself is not influencing development. Age ranges for the testing cohort should be limited to 2 to 4 weeks with defined windows for "pre-wean" (<3 weeks), "young" (4 to 7 weeks of age), "adult" (≥8 weeks of age), and "aged" (≥20 weeks of age). Litters should be separated by sex by 4 weeks of age to avoid unintended brother-sister matings.

3. Sample sizes should be adequate and informed by historical data based on the minimum sample size required to achieve statistical significance or statistically through power calculations. Depending on the behavioral test, typical sample sizes may range from n = 10 to 20 per sex, per genotype/treatment group. Considerations may also need to estimate additional n size for potential attrition rates in the case of chronic studies or when morbidity or mortality may be associated with the animal model.

4. When ordering from a commercial vendor, it is important to record not only the vendor and the strain and substrain information (i.e., C57BL/6J, C57BL/6NJ, C57BL/6NTac) but also the location the animals were reared including the specific colony room, details of the specific diet (i.e., diet vendor, dietary components, % fat), water regimen (i.e., tap, filtered, acidified), and any other environmental parameters (e.g., background music in the rearing environment). As already stated, it is important to not only be aware of behavioral differences across substrains but also within strains; even within a strain divergent behaviors may occur as a result of variations in rearing environment which could be related to diet, microbiome, or other unknown factors. Therefore, it is best practice that control mice are littermate controls or at best that the control strain be maintained from a consistent colony room from the same vendor. It should be noted that commercial vendors not only have multiple production facilities across the world but also have, even within the same location, variations on husbandry practices (e.g., high versus low barrier rooms) that can influence behavior.

5. Mice obtained either from a commercial vendor or generated from another laboratory should be acclimated to the laboratory environment both to recover from shipping stress as well as to ensure they become entrained to the new laboratory environment, which has been reported to take at least 5 days (Obernier and Baldwin, 2006).

6. Animals should be housed for validation studies similar to the manner in which future experimental studies will be carried out in order to establish consistent operating procedures. Behavioral responses can vary with housing density. Mice housed together in groups or pairs establish dominance hierarchies that may also contribute to variability across certain behavioral assays (e.g., social behavior). In some strains, excessive fighting, particularly in males, may require the need to separate subjects from the initiation of a testing battery so as to minimize having to remove them or treat them during a study. Assay validation under different housing densities may be required to understand whether housing density is a contributing variable to the behavior. It has been reported that housing density influences social and anxiety-like behaviors dependent upon whether subjects are housed individually, group housed within mixed genotypes, or housed within same genotype groups (Yang et al., 2011).

7. Genotypes should be blinded to minimize any potential for bias. To ensure randomization and counterbalancing, subject identifications can be associated with "A," "B," or "C" as coded genotype, and blinding should be maintained until data analysis has been completed.

8. Considerations should be made for handling procedures and cage changes prior to behavioral testing. In general, behavioral experiments should not be conducted on the day of a cage change. Depending on the behavioral test (e.g, social or repetitive behavioral assessments), a minimum 3-day interval between a cage change and the behavioral test is recommended. During tests that require daily assessments such as the Morris Water Maze, cage changing over the multi-day assay increases variability. Importantly, the testing battery should include scheduled cage changes so that the data can be reproduced. Further, the manner in which mice are handled for cage changes by husbandry staff should be consistent, and the use of tail forceps is not preferred.

9. Mice will require unique identifications so as to identify each individual. Several methods have been used for permanent identification of mice including ear tags, ear notching, tail tattooing, digit notching, RFID implants, and tail labeling with nontoxic marker. Each have their own benefits and limitations. Digit notching is not preferred as this could impact motor activity and cannot be consistent across group housed animals. Although ear tags are easily readable, they can be ripped out requiring treatment. Tail tattooing and subcutaneous implants have risk, although minimal of inflammatory responses which may or may not impact behavioral outcomes. Ear notching is the most common, although it requires training. Regardless of the method used for permanent identification, any of these procedures should be done prior to the start of behavioral testing, leaving a sufficient amount of recovery period (e.g., 1 week). Further for consistency, all subjects within a cohort should receive the same type of identification and the identical handling procedure.

10. For compound testing experiments, test subjects should be drug-naïve mice so as to avoid any drug tolerance that may occur with repeated dosing.

Environmental Considerations

1. Temperature, humidity, type of bedding, and lighting levels in the housing room and testing rooms should be in accordance with IACUC. Variables such as environmental enrichment and background music in housing rooms should be described. Presence of and type of enrichment (i.e., foraging and nesting materials, huts, tubes) is a variable and should be made consistent across the cohort of test subjects.

2. Details of the light:dark cycle and time of day with respect to lights on:off, whether testing was conducted under an inverted light cycle, or whether testing was conducted under red lights should be reported. Behavior can vary with respect to time of day with higher activity levels closer to or during the dark cycle. Counterbalancing representative samples from each treatment/sex/genotype may be required across multiple days if the experiment requires an extended amount of time to complete. Importantly, subjects of a single group should not all be tested at the same time but rather should be randomized and counterbalanced across the entire testing period.

3. Lighting levels should be standardized for individual tests dependent on the specific environmental requirements of the behavioral test and for consistency; levels should be detailed as part of the testing protocol and published methods. Commercially available lux meters to measure lighting levels and decibel meters to measure sound levels can be readily purchased. A lux meter should be used to measure lighting in the testing environment and, once established, should be maintained for each assay as part of its protocol. Ambiguous terms such as "dim" or "high" should not be used; instead specific lux/lumen levels should be reported as well as the type of lighting (i.e., fluorescent, LED). If possible during facility design, dimmable lighting with on/off controls that are not directly overhead is preferred. Direct ceiling lighting often contributes to glare issues when automated tracking or video recording is being used.

4. Background noise should be recorded with a commercially available sound meter and maintained consistent as part of the testing protocol. Commercially available white noise generators can be purchased to maintain a consistent background noise level which helps to eliminate disruptions from random noise in adjacent testing areas. Typical background noise levels of <70 dB are appropriate.

5. Disruptions to the housing environment should be minimized, and any disruption to the environment during testing should be recorded. Phones should be silenced, and appropriate "Do Not Enter" signage should be posted on procedure room doors to eliminate unnecessary disruptions. Additional engineering controls, including minimizing clicking of doors as they are opened or closed and ticking of the secondhand of clocks, should be eliminated so as not to serve as disruptive audio stimuli during sensitive behavioral testing. Only timers and stopwatches with silence fea-tures should be used when required during the test. Alternatively, the audio feature on most stopwatches and timers can be easily removed (Yang et al., 2011).

Behavioral Testing and Tracking Equipment

1. Behavioral testing equipment can be purchased from a variety of commercial vendors or can be fabricated by skilled craftsman. Details of equipment dimensions and vendor information should be reported inclusive of the material used for fabrication and color. This includes detailing "custom made" objects or visual cues for which dimensions and photos should be provided in the methods section of the protocol and may be critical to reproducing the data. Selection of

equipment and fabrication material is an important consideration. While acrylic is an inexpensive material that is often used to fabricate behavioral testing arenas, it is highly susceptible to scratches and cracking due to repeated exposure of concentrated ethanol solution which is typically used as the sanitizing agent between subjects; therefore polycarbonate is a better alternative. When mice of varying coat colors are being tested, automated or video tracking may necessitate a change of background color for contrast. To minimize this, infrared reflecting background systems paired with infrared cameras can be purchased that eliminate the additional variable of background color and also minimize video glare.

2. Behavioral tracking software can be purchased from any number of vendors that specialize in mouse behavior tracking. While in theory behavioral tracking software should help facilitate experiments and minimize the additional stimulus of the experimenter in close proximity to the test subjects, it is important to ensure that the behavior being evaluated is what is precisely being captured by the software. For example, various background strains of mice vary in their swimming behavior in the forced swim test. Specifically, obese mice, although not actively swimming, tend to bob from side to side and these data might be calculated as swimming time on an automated system that was calibrated for swimming and immobility behaviors in a non-obese strain (e.g., C57BL/6J). Therefore, time and resources should be allocated to calibrate the automated tracking software, ensuring it is capturing the desired behavior precisely, in line with direct observations from more than one trained observer. Adjustments to the tracking software can then be employed as required, if such capabilities are available. It is suggested that a correlation or statistical proof of accuracy be included in the laboratory’s first publication of the automated measurement for validation.

3. Visual cues may be required for certain tests or may need to be eliminated to avoid extra-maze or unintended visual cues including the visual presence of the experimenter. The selection of specific visual stimuli for certain behavioral tasks (i.e., recognition memory) may require pretesting to ensure the cues are salient and that there is not a bias or preference for one cue over another. In studies of recognition memory where multiple visual cues are used, cue bias can be minimized by counterbalancing the presentation of the visual cues across subjects within a treatment group (i.e., odd numbered subjects are assigned cue A as correct while even numbered subjects are assigned cue B as correct). To minimize extra-maze visual cues, a curtain can be used to surround the perimeter of the testing equipment which would also eliminate the visual presence of the experimenter moving around the testing room during the observation period.

Test Compounds

1. Drugs should be procured in powder form instead of prepackaged solutions for which the vehicle constituents may be undisclosed. For example, diazepam is a standard anxiolytic-agent that can be used to demonstrate anxiolytic-like effects

in mice. Use of the clinical compound Valium in its prepackaged liquid form, however, is in a formulation of undisclosed constituents. Therefore it is recommended that research grade diazepam powder be procured and formulated with known and behaviorally acceptable vehicles. In general, all compounds should be formulated fresh daily unless storage conditions in the absence of preservatives are well established, and, importantly, caution should be taken in preparing compounds with unknown or high instability (i.e., peptides) as even mild vortexing can degrade the compound.

2. Drug concentrations should be calculated as the active compound relative to the percentage of the molecule that is inactive or a salt molecule (% active moiety) and accurately reported (Sukoff Rizzo, 2016). Details of how the drug was formulated—including how much drug was weighed, the volume of the diluents, requirements for heating (and the specific temperature), and vortexing or mixing requirements—are all important details that should be recorded, as well as whether the drug was in solution or dosed as a suspension and the pH of the concentration dosed including any requirements for titrating with acid or base to facilitate solubility.

3. Many reagents used as excipients for formulating test compounds in vehicle may produce unexpected behavioral effects even in the absence of the test compound, depending on concentration used. Considerations should be made for pretesting a novel vehicle for unexpected behavioral effects prior to testing the drug (Castro et al., 1995; Lin et al., 1998; Colucci et al., 2008; Rivers-Auty and Ashton, 2013).

4. Prior to initiating dosing, subjects should be randomized and preassigned to a treatment group with careful attention to counterbalancing treatment groups across test observations and days. A second technician familiar with the study but not responsible for conducting the observations should assist with the blinding of the drug vials. Blinding of drug vials can be facilitated similar to the simple letter coding used to blind for genotype (e.g., “A,” “B,” “C,” “D”).

METHODS

Acclimation to the Testing Environment

1. Test room conditions should be set prior to introducing the subjects into the testing environment for the day (see Environmental Considerations). If space permits, an ante room for acclimation and a separate room for dosing are optimal. Both areas are separate from the behavioral testing room, but adjacent such that they do not require extensive moving of the mice for dosing and testing. Tests that use audio cues, for example, require acclimation in an anteroom or space outside but immediately adjacent to the testing room in order to prevent the test subjects from acclimating to the test stimuli prior to testing.

2. Prior to the initiation of the acclimation period, mice should be briefly handled (to ensure they are healthy prior to being tested) and weighed if the procedure

requires body weight or if dosing will be conducted. Body weights should be recorded for individual mice prior to dosing and should not be recorded as a mean for the cage or recorded the night before the test.

3. Test subjects should be left undisturbed for 60 min prior to testing in order to acclimate to the testing environment.

Experiment Timing

1. A spreadsheet which maps out the precise timing of the experimental events of the test should be created before the start of each test which identifies the precise dose time, the start and end time of each subject's observation, the time to clean the equipment between subjects, and the dose time and evaluation times of the next sequential test subjects. A notes section should also be provided to add important documentation that should be noted (i.e., bad injection, noise disruption during trial) which can help identify any spurious data during the analysis. The pretreatment time for individual animals should be carefully planned as well as the timing required to clean the testing environment between subjects. Dosing of a test subject should be avoided during the recording of another test subject so as to avoid altering behavior in response to disruptions induced by movement of cages or vocalizations associated with restraint.

2. Separate needles and syringes should be used for each test subject. Labeling and prefilling of syringes ahead of the start of the experiment are recommended to minimize errors during the actual testing period when a high level of multitasking is required (i.e., dosing, cleaning, observing). However, if the drug is a suspension, the syringes should not be prefilled as settling could occur. In these cases, it may be important for the drug to remain on a stir plate to maintain the suspension, and syringes should not be filled until immediately prior to dosing.

Behavioral Testing

1. How the animal is introduced into the testing apparatus is a critical detail of the experiment and should be documented as part of the protocol. The test subject should consistently be placed in a similar location within the test apparatus (i.e., "facing the center"), and this should be noted as part of the standard procedure.

2. During the observation period, all distractions should be minimized, and any disruptions should be recorded on a run sheet. If more than one mouse is being tested simultaneously in adjacent equipment, the technician should wait until the session has completed for all test subjects prior to removing individual subjects so as to avoid any noise disruptions from equipment movement.

3. The method of how the test subject is treated at the completion of the observation should be recorded. At the conclusion of the test, mice can either be returned to their home cage or an alternate holding cage until all cage mates have completed the task. Considerations should be made for returning test subjects back to a group setting at the conclusion of testing, or rather test subjects may be placed in

a separate holding cage until all cage mates have been completed testing. For example, in a group housed setting when treatments are randomized within a cage and a drug induces a behavioral effect (e.g., hyperactivity) which could result in arousal of their acclimating cage mates, then an alternate post-testing holding arrangement should be considered as may also be considered during the pretreatment period.

4. Between test subjects, urine and fecal boli should be removed from the testing apparatus, and a sanitizing agent that minimizes odors (e.g., 70% ethanol) should be generously applied to eliminate scent cues. The testing apparatus should be allowed to dry prior to placing the next subject into the apparatus.

DATA ANALYSIS

1. Data should be presented and analyzed as raw values and illustrated in graph form with the distribution of the data points if feasible.
2. An appropriate statistical analysis should be chosen based on the data generated. Typically an ANOVA is used to analyze the data for drug screening studies for a single behavioral endpoint with multiple doses (dose-response curve) with an appropriate post hoc test (e.g., Dunnett's post hoc test with vehicle as control). Genotype × drug × sex usually needs a Tukey test for highest stringency.
3. Assessment of multiple time points may require a two-way repeated measures ANOVA as treatment × time (e.g., activity in the open field recorded as 5 min time bins over the course of an hour). If multiple data points are being generated and analyzed, then the analysis and presentation of the data should reflect the multiple points (i.e., time course as opposed to a single time point).
4. The technician should only be unblinded after the data have been analyzed.
5. Negative findings should be disclosed in addition to positive data.

SUMMARY

Behavioral phenotyping and psychopharmacology are sciences that require highly specialized levels of training. The experimental rigor and standards are beyond those of most simple biological wet lab bench assays. The present article provides a pragmatic outline for establishing behavioral phenotyping and testing tailored behavioral batteries in the laboratory as well as guidelines for the training of investigators and technical staff. Our article highlights the previously undescribed and under-reported specific details that are fundamental to execute behavioral assays in a manner that is sensitive to detect subtle behavioral changes. Moreover, these methods, if conducted properly, should yield reproducibility and reliability both in intra- and interlaboratory environments.

ACKNOWLEDGMENTS

This work is supported by The Jackson Laboratory's Center for Biometric Analysis and The MIND Institute's Intellectual and Developmental Disabilities Research Center (NICHD-IDDRC HDHD079125).

LITERATURE CITED

Bourin M, Poncelet M, Chermat R, Simon P. The value of the reserpine test in psychopharmacology Arzneimittelforschung. 1983; 33:1173–1176. [PubMed: 6685496]

Brown JW, Rueter LE, Zhang M. Predictive validity of a MK-801-induced cognitive impairment model in mice: Implications on the potential limitations and challenges of modeling cognitive impairment associated with schizophrenia preclinically. Prog. Neuropsychopharmacol Biol. Psychiatry. 2014; 49:53–62. . DOI: 10.1016/j.pnpbp.2013.11.008

Bryant CD, Zhang NN, Sokoloff G, Fanselow MS, Ennes HS, Palmer AA, McRoberts JA. Behavioral Differences among C57BL/6 Substrains: Implications for Transgenic and Knockout Studies J. Neurogenet. 2008; 22:315–331. . DOI: 10.1080/01677060802357388 [PubMed: 19085272]

Buccafusco JJ, Methods of Behavior Analysis in Neuroscience 2nd. CRC Press; Boca Raton, Fla: 2009

Castagné V, Porsolt RD, Moser P. Use of latency to immobility improves detection of antidepressant-like activity in the behavioral despair test in the mouse Eur J. Pharmacol. 2009; 616:128–133. . DOI: 10.1016/j.ejphar.2009.06.018 [PubMed: 19549518]

Castro CA, Hogan JB, Benson KA, She-hata CW, Landauer MR. Behavioral effects of vehicles: DMSO, ethanol, Tween-20, Tween-80, and emulphor-620 Phar-macol. Biochem. Behav. 1995; 50:521–526. . DOI: 10.1016/0091-3057(94)00331-9

Colucci M, Maione F, Bonito MC, Piscopo A, Di Giannuario A, Pieretti S. New insights of dimethyl sulphoxide effects (DMSO) on experimental in vivo models of nociception and inflammation Pharmacol. Res. 2008; 57:419–425. . DOI: 10.1016/j.phrs.2008.04.004 [PubMed: 18508278]

Crabbe JC, Wahlsten D, Dudek BC. Genetics of Mouse Behavior: Interactions with Laboratory Environment Science. 1999; 284:1670–1672. . DOI: 10.1126/science.284.5420.1670 [PubMed: 10356397]

Crabbe JC, Metten P, Cameron AJ, Wahlsten D. An analysis of the genetics of alcohol intoxication in inbred mice Neurosci. Biobehav. Rev. 2005; 28:785–802. . DOI: 10.1016/j.neubiorev.2004.08.002 [PubMed: 15642621]

Crawley JN, Behavioral Phenotyping of Transgenic and Knockout Mice 2nd. What's Wrong with My Mouse?Wiley-Liss; New York: 2007

Crawley JN, Paylor R. A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice Horm. Behav. 1997; 31:197–211. . DOI: 10.1006/hbeh.1997.1382 [PubMed: 9213134]

Crawley JN, Belknap JK, Collins A, Crabbe JC, Frankel W, Henderson N, Hitzemann RJ, Maxson SC, Miner LL, Silva AJ, Wehner JM, Wynshaw-Boris A, Pay-lor R. Behavioral phenotypes of inbred mouse strains: Implications and recommendations for molecular studies Psychopharmacology. 1997; 132:107–124. . DOI: 10.1007/s002130050327 [PubMed: 9266608]

Cryan JF, Dinan TG. Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour Nat. Rev. Neurosci. 2012; 13:701–712. . DOI: 10.1038/nrn3346 [PubMed: 22968153]

Cryan JF, Mombereau C. In search of a depressed mouse: Utility of models for studying depression-related behavior in genetically modified mice Mol. Psychiatry. 2004; 9:326–357. . DOI: 10.1038/sj.mp.4001457 [PubMed: 14743184]

Cryan JF, Mombereau C, Vassout A. The tail suspension test as a model for assessing antidepressant activity: Review of pharmacological and genetic studies in mice Neurosci. Biobehav. Rev. 2005; 29:571–625. . DOI: 10.1016/j.neubiorev.2005.03.009 [PubMed: 15890404]

Darmani NA, Martin BR, Pandey U, Glen-non RA. Do functional relationships exist between 5-HT1A and 5-HT2 receptors? Pharmacol. Biochem. Behav. 1990; 36:901–906. . DOI: 10.1016/0091-3057(90)90098-3 [PubMed: 2145593]

De Aceto MD, Mackean DB, Pearl J. Effects of opiates and opiate antagonists on the straub tail reaction in mice Br. J. Pharmacol. 1969; 36:255. . doi: 10.1111/j.1476-5381.1969.tb09500.x

Ding M, Turner AJ, Ramkumar V, Hughes LF, Trammell RA, Toth LA. Lack of association of a spontaneous mutation of the Chrm2 gene with behavioral and physiologic phenotypic differences in inbred mice Comp. Med. 2010; 60:272–281. [PubMed: 20819376]

Dunn AJ, Swiergiel AH. Effects of interleukin-1 and endotoxin in the forced swim and tail suspension tests in mice Pharmacol. Biochem. Behav. 2005; 81:688–693. . DOI: 10.1016/j.pbb.2005.04.019 [PubMed: 15982728]

Dyer RS, Weldon DA. Blindness-induced hyperactivity in several strains of mice Physiol. Behav. 1975; 15:439–441. . DOI: 10.1016/0031-9384(75)90211-5 [PubMed: 1221451]

Everitt JI, Foster PMD. Laboratory animal science issues in the design and conduct of studies with endocrine-active compounds ILAR J. 2004; 45:417–424. . DOI: 10.1093/ilar.45.4.417 [PubMed: 15454680]

Fuchs H, Gailus-Durner V, Adler T, Aguilar-Pimentel JA, Becker L, Calzada-Wack J, Da Silva-Buttkus P, Neff F, Götz A, Hans W, Holter SM, Hörsch M, Kastenmüller G, Kemter E, Lengger C, Maier H, Matloka M, Möller G, Naton B, Prehn C, Puk O, Rácz I, Rathkolb B, Römisch-Margl W, Rozman J, Wang-Sattler R, Schrewe A, Stöger C, Tost M, Adamski J, Aigner B, Beckers J, Behrendt H, Busch DH, Esposito I, Graw J, Illig T, Ivandic B, Klingenspor M, Klopstock T, Kremmer E, Mempel M, Neschen S, Ollert M, Schulz H, Suhre K, Wolf E, Wurst W, Zimmer A, Hrabě de Angelis M. Mouse phenotyping Methods. 2011; 53:120–135. . DOI: 10.1016/j.ymeth.2010.08.006 [PubMed: 20708688]

Griebel G, Misslin R, Pawlowski M, Vogel E. Meta-chlorophenylpiperazine enhances neophobic and anxious behavior in mice Neuroreport. 1991; 2:627–629. . DOI: 10.1097/00001756-199110000-00019 [PubMed: 1756245]

Griebel G, Belzung C, Perrault G, Sanger DJ. Differences in anxiety-related behaviours and in sensitivity to diazepam in inbred and outbred strains of mice Psychopharmacology. 2000; 148:164–170. . DOI: 10.1007/s002130050038 [PubMed: 10663431]

Haberzettl R, Fink H, Bert B. Role of 5-HT(1A)- and 5-HT(2A) receptors for the murine model of the serotonin syndrome J. Pharmacol. Toxicol. Methods. 2014; 70:129–133. . DOI: 10.1016/j.vascn.2014.07.003 [PubMed: 25087754]

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research PLoS Biol. 2010; 8:e1000412. . doi: 10.1371/journal.pbio.1000412 [PubMed: 20613859]

Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals PLoS ONE. 2009; 4:e7824. . doi: 10.1371/journal.pone.0007824 [PubMed: 19956596]

Klinkenberg I, Blokland A. The validity of scopolamine as a pharmacological model for cognitive impairment: A review of animal behavioral studies. *Neurosci* Biobehav. Rev. 2010; 34:1307–1350. . DOI: 10.1016/j.neubiorev.2010.04.001

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD. A call for transparent reporting to optimize the predictive value of preclinical research Nature. 2012; 490:187–191. . DOI: 10.1038/nature11556 [PubMed: 23060188]

Lin HQ, Burden PM, Johnston GAR. Propylene glycol elicits anxiolytic-like responses to the elevated plus-maze in male mice J. Pharm. Pharmacol. 1998; 50:1127–1131. . DOI: 10.1111/j.2042-7158.1998.tb03323.x [PubMed: 9821659]

Mandillo S, Tucci V, Holter SM, Meziane H, Banchaabouchi MA, Kallnik M, Lad HV, Nolan PM, Ouagazzal AM, Coghill EL, Gale K, Golini E, Jacquot S, Krezel W, Parker A, Riet F, Schneider I, Marazziti D, Auwerx J, Brown SD, Chambon P, Rosenthal N, Tocchini-Valentini G, Wurst W. Reliability, robustness, and reproducibility in mouse behavioral phenotyping: A cross-laboratory study Physiol. Genomics. 2008; 34:243–255. . DOI: 10.1152/physiolgenomics.90207.2008 [PubMed: 18505770]

McIlwain KL, Merriweather MY, Yuva-Paylor LA, Paylor R. The use of behavioral test batteries: Effects of training history Physiol. Behav. 2001; 73:705–717. . DOI: 10.1016/S0031-9384(01)00528-5 [PubMed: 11566205]

Meisenberg G, Simmons WH. Behavioral effects of intracerebroventricularly administered neurohypophyseal hormone analogs in mice Pharmacol. Biochem. Behav. 1982; 16:819–825. . DOI: 10.1016/0091-3057(82)90242-8 [PubMed: 7089039]

Mogil JS, Kest B, Sadowski B, Belknap JK. Differential genetic mediation of sensitivity to morphine in genetic models of opiate antinociception: Influence of nociceptive assay J. Pharmacol. Exp. Ther. 1996; 276(2):532–544. [PubMed: 8632319]

Moy SS, Nonneman RJ, Shafer GO, Nikolova VD, Riddick NV, Agster KL, Baker LK, Knapp DJ. Disruption of social approach by MK-801, amphetamine, and fluoxetine in adolescent C57BL/6J mice Neurotoxicol. Teratol. 2013; 36:36–46. . DOI: 10.1016/j.ntt.2012.07.007 [PubMed: 22898204]

Nath C, Gupta MB, Pantaik GK, Dhawan KN. Morphine-induced straub tail response: Mediated by central μ2-opioid receptor Eur J. Pharmacol. 1994; 263:203–205. DOI: 10.1016/0014-2999(94)90543-6. [PubMed: 7821354]

Obernier JA, Baldwin RL. Establishing an appropriate period of acclimatization following transportation of laboratory animals ILAR J. 2006; 47:364–369. . DOI: 10.1093/ilar.47.4.364 [PubMed: 16963816]

Olsson IA, Dahlborn K. Improving housing conditions for laboratory mice: A review of “environmental enrichment” Lab. Anim. 2002; 36:243–270. . DOI: 10.1258/002367702320162379 [PubMed: 12144738]

Oswald S, Balice-Gordon R. Rigor or mortis: Best practices for preclinical research in neuroscience Neuron. 2014; 84:572–581. . DOI: 10.1016/j.neuron.2014.10.042 [PubMed: 25442936]

Paylor R, Spencer CM, Yuva-Paylor LA, Pieke-Dahl S. The use of behavioral test batteries, II: Effect of test interval Physiol. Behav. 2006; 87:95–102. . DOI: 10.1016/j.physbeh.2005.09.002 [PubMed: 16197969]

Ralph RJ, Paulus MP, Geyer MA. Strain-specific effects of amphetamine on prepulse inhibition and patterns of locomotor behavior in mice J. Pharmacol. Exp. Ther. 2001; 298:148–155. [PubMed: 11408536]

Rasmussen S, Glickman G, Norinsky R, Quimby FW, Tolwani RJ. Construction noise decreases reproductive efficiency in mice J. Am. Assoc. Lab. Anim. Sci. 2009; 48:363–370. [PubMed: 19653943]

Rivers-Auty J, Ashton JC. Vehicles for lipophilic drugs: Implications for experimental design, neuroprotection, and drug discovery Curr. Neurovasc. Res. 2013; 10:356–360. . DOI: 10.2174/15672026113109990021 [PubMed: 23937198]

Rizzo SJ, Edgerton JR, Hughes ZA, Brandon NJ. Future viable models of psychiatry drug discovery in pharma J. Biomol. Screen. 2013; 18:509–521. . DOI: 10.1177/1087057113475871 [PubMed: 23392517]

Rodgers RJ, Cole JC, Aboualfa K, Stephenson LH. Ethopharmacological analysis of the effects of putative ‘anxiogenic’ agents in the mouse elevated plus-maze Pharmacol. Biochem. Behav. 1995; 52:805–813. . DOI: 10.1016/0091-3057(95)00190-8 [PubMed: 8587923]

Schellinck HM, Cyr DP, Brown RE. How many ways can mouse behavioral experiments go wrong? Confounding variables in mouse models of neurodegenerative diseases and how to control them Adv. Study Behav. 2010; 41:255–366. . DOI: 10.1016/S0065-3454(10)41007-4

Schneider I, Tirsch WS, Faus-Kebler T, Becker L, Kling E, Austin Busse RL, Bender A, Feddersen B, Tritschler J, Fuchs H, Gailus-Durner V, Englmeier K-H, Hrabé de Angelis M, Klopstock T. Systematic, standardized and comprehensive neurological phenotyping of inbred mice strains in the German Mouse Clinic J. Neurosci. Method. 2006; 157:82–90. . DOI: 10.1016/j.jneumeth.2006.04.002

Shepherd JK, Grewal SS, Fletcher A, Bill DJ, Dourish CT. Behavioural and pharmacological characterisation of the elevated “zero-maze” as an animal model of anxiety Psychopharmacology. 1994; 116:56–64. . DOI: 10.1007/BF02244871 [PubMed: 7862931]

Silverman JL, Crawley JN. The promising trajectory of autism therapeutics discovery Drug. Discov. Today. 2014; 19:838–844. . DOI: 10.1016/j.drudis.2013.12.007 [PubMed: 24362109]

Silverman JL, Smith DG, Rizzo SJ, Karras MN, Turner SM, Tolu SS, Bryce DK, Smith DL, Fonseca K, Ring RH, Crawley JN. Negative allosteric modulation of the mGluR5 receptor reduces repetitive behaviors and rescues social deficits in mouse models of autism Sci. Transl. Med. 2012; 4:131ra51. . doi: 10.1126/scitranslmed.3003501

Simon MM, Greenaway S, White JK, Fuchs H, Gailus-Durner V, Wells S, Sorg T, Wong K, Bedu E, Cartwright EJ, Dacquin R, Djebali S, Estabel J, Graw J, Ingham NJ, Jackson IJ, Lengeling A, Mandillo S, Marvel J, Meziane H, Preitner F, Puk O, Roux M, Adams DJ, Atkins S, Ayadi A, Becker L, Blake A, Brooker D, Cater H, Champy MF, Combe R, Danecek P, di Fenza A, Gates H, Gerdin AK, Golini E, Hancock JM, Hans W, Hölter SM, Hough T, Jurdic P, Keane TM, Morgan H, Müller W, Neff F, Nicholson G, Pasche B, Roberson LA, Rozman J, Sanderson M, Santos L, Selloum M, Shannon C, Southwell A, Tocchini-Valentini GP, Vancollie VE, Westerberg H, Wurst W, Zi M, Yalcin B, Ramirez-Solis R, Steel KP, Mallon AM, de Angelis MH, Herault Y, Brown SD. A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains Genome Biol. 2013; 14:R82. . doi: 10.1186/gb-2013-14-7-r82 [PubMed: 23902802]

Sukoff Rizzo SJ, Proetzel G, Wiles M. Mouse Models for Drug Discovery Repetitive behavioral assessments for compound screening in mouse models of autism spectrum disordersSpringer; New York: 2016 In press.

Sukoff Rizzo SJ, Neal SJ, Hughes ZA, Beyna M, Rosenzweig-Lipson S, Moss SJ, Brandon NJ. Evidence for sustained elevation of IL-6 in the CNS as a key contributor of depressive-like phenotypes Transl. Psychiatry. 2012; 2:e199. . doi: 10.1038/tp.2012.120

Tricklebank MD, and Garner JP, 2012 The possibilities and limitations of animal models for psychiatric disorders In Drug Discovery for Psychiatric Disorders (Rankovic Z; , Bingham M, Nestler EJ, , and Hargreaves R. , eds.) pp. 534–556 . The Royal Society of Chemistry , London .

Turner CA, Lewis MH, King MA. Environmental enrichment: Effects on stereotyped behavior and dendritic morphology Dev. Psychobiol. 2003; 43:20–27. . DOI: 10.1002/dev.10116 [PubMed: 12794775]

Unger EF. All is not well in the world of translational research J. Am. Coll. Cardiol. 2008; 50:738–740. . DOI: 10.1016/j.jacc.2007.04.067

van der Staay FJ, Steckler T. The fallacy of behavioral phenotyping without standardisation Genes Brain Behav. 2002; 1:9–13. . DOI: 10.1046/j.1601-1848.2001.00007.x [PubMed: 12886945]

Varty GB, Walters N, Cohen-Williams M, Carey GJ. Comparison of apomorphine, amphetamine and dizocilpine disruptions of prepulse inhibition in inbred and outbred mice strains Eur. J. Pharmacol. 2001; 424:27–36. . DOI: 10.1016/S00142999(01)01115-3 [PubMed: 11470257]

Wahlsten D, 1st. MouseBehavioral Testing: How to UseMice in Behavioral NeuroscienceAcademic Press; San Diego: 2010

Wahlsten D, Metten P, Phillips TJ, Boehm SL, Burkhart-Kasch S, Dorow J, Doerksen S, Downing C, Fogarty J, Rodd-Henricks K, Hen R, McKinnon CS, Merrill CM, Nolte C, Schalomon M, Schlumbohm JP, Sibert JR, Wenger CD, Dudek BC, Crabbe JC. Different data from different labs: Lessons from studies of geneenvironment interaction J. Neurobiol. 2003; 54:283–311. . DOI: 10.1002/neu.10173 [PubMed: 12486710]

Würbel H. Behaviour and the standardization fallacy Nature Genet. 2000; 26:263. . doi: 10.1038/81541 [PubMed: 11062457]

Würbel H. Behavioural phenotyping enhanced: Beyond (environmental) standardization Genes Brain Behav. 2002; 1:3–8. . DOI: 10.1046/j.1601-1848.2001.00006.x [PubMed: 12886944]

Yamada J, Sugimoto Y, Horisaka K. The behavioural effects of 8-hydroxy-2-(di-n-propylamino) tetralin (8-OH-DPAT) in mice Eur. J. Pharmacol. 1988; 154:299–304. . DOI: 10.1016/0014-2999(88)90205-1 [PubMed: 2976671]

Yang M, Silverman JL, Crawley JN. Automated three-chambered social approach task for mice Curr. Protoc. Neurosci. 2011; 56:8.26.1–8.26.16. . DOI: 10.1002/0471142301.ns0826s56

Yang M, Perry K, Weber MD, Katz AM, Crawley JN. Social peers rescue autism-relevant sociability deficits in adolescent mice Autism Res. 2011; 4:17–27. . DOI: 10.1002/aur.163 [PubMed: 20928844]

Yonekawa WD, Kupferberg HJ, Woodbury DM. Relationship between pentylenetetrazol-induced seizures and brain pentylenetetrazol levels in mice J. Pharmacol. Exp. Ther. 1980; 214:589–593. [PubMed: 7400961]

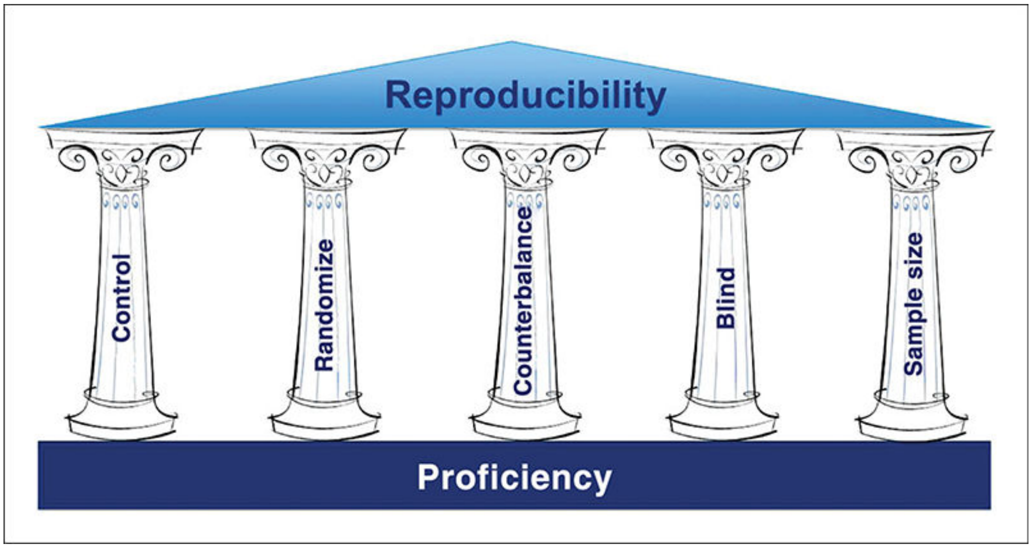

Figure 1.

The pillars of reproducibility. The overarching goal of reproducibility can be achieved by developing a rigorous experimental design that takes into consideration the applications of blinding, randomization, counterbalancing, suitable sample sizes, and the inclusion of appropriate controls. It is supported by highly trained technical staff with demonstrated proficiency at conducting the assays.

Table 1

Commercially Available Reagents That Can Be Used to Demonstrate Assay Validation and Technician Proficiency for Behavioral Testing and Identification of Abnormal Behaviors in Mice

Behavior	Commercially available positive control reagents	Assay(s)	References
Antidepressant-like effects	Fluoxetine, imipramine, desipramine	Forced swim test; tail suspension test	Cryan et al. (2005); Castagne et al. (2009)
Depressive-like effects	LPS[a], IL-6; mouse models: MRL, LPR	Forced swim test; tail suspension test	Dunn and Swiergiel (2005); Sukoff Rizzo et al. (2012)
Anxiolytic-like effects	Diazepam, alprazolam, chlordiazepoxide	Elevated plus maze, elevated zero maze, stress-induced hyperthermia, light-dark, 4-plate conflict assay (Vogel, Geller-Seifter)	Shepherd et al. (1994); Griebel et al. (2000)
Anxiogenic-like effects	FG7142, mCPP[b]; mouse strains: BALBc/J relative to C57BL/6J	Elevated plus maze, elevated zero maze	Griebel et al. (1991); Shepherd et al. (1994); Rodgers et al. (1995); Crawley et al. (1997)
Nociceptive behavior	Morphine	Hot plate, Von Frey	Mogil et al. (1996)
Social behavior deficits	Amphetamine, MK-801; mouse strains: BTBR relative to C57BL/6J	3-chamber social approach, reciprocal social interaction	Silverman et al. (2012); Moy et al. (2013)
Induction of repetitive behaviors	Oxytocin; BTBR mice	Repetitive grooming	Meisenberg and Simmons (1982); Silverman et al. (2012)
Repetitive jumping behavior	C58/J mice	Repetitive jumping	Silverman et al. (2012)
Motor alterations (hypoactivity, ataxia)	Ethanol	Rotarod, gait, grip strength	Crabbe et al. (2005)
Cognitive impairment	Scopolamine, MK-801	Water maze, novel object recognition, novel spatial recognition, spontaneous alternation, hole board learning, contextual fear conditioning, operant/ touchscreen tasks	Klinkenberg, and Blokland (2010); Brown et al. (2014)
Pre-pulse inhibition	Amphetamine, MK-801, apomorphine; mouse strains: C57BL/6J relative to C57BL/6N	Pre-pulse inhibition of acoustic startle	Varty et al. (2001); Simon et al. (2013)
Hyperactivity	Amphetamine	Open field	Ralph et al. (2001); Varty et al. (2001)
Head twitch	DOI[c]	SHIRPA, Irwin Screen	Darmani et al. (1990)
Straub tail	Morphine; 8-OH-DPAT	SHIRPA, Irwin Screen	Nath et al. (1994); Yamada et al. (1988)
Tremor	oxotremorine	SHIRPA, Irwin Screen	Ding et al. (2010)
Piloerection	LPS[a]	SHIRPA, Irwin Screen	Dunn and Swiergiel (2005)
Seizure	*PTZ*[d]	SHIRPA, Irwin Screen	Yonekawa et al. (1980)
Forepaw treading	8-OH-DPAT	SHIRPA, Irwin Screen	Yamada et al. (1988)
Ptosis	Reserpine	SHIRPA, Irwin Screen	Bourin et al. (1983)
Serotonin syndrome	8-OH-DPAT	Serotonin syndrome, Irwin Screen	De Aceto et al. (1969); Yamada et al. (1988)

[a] LPS = lipopolysaccharide.

[b] mCCP = meta-chlorophenylpiperazine

[c] DOI = 1-(2, 5-dimethoxy-4-iodophenyl)-2-aminopropane.

[d] PTZ = pentylenetetrazol.

Table 2

Potential Confounds Associated with Behavioral Assays and Methods for Assuring Appropriate Interpretation

Assay	Confounding/masking behavior	Assurance of phenotype
Contextual fear conditioning and other assays requiring shock stimuli	Hyperactivity, insensitivity to shock level (analgesia)	Titrate shock levels for independent genotype. Check activity levels (hyperactivity).
Repetitive behaviors (i.e., grooming, marble burying test)	Hyperactivity, sedation, hypoactivity, ataxia	Confirm lack of motor alterations in alternative test (e.g, open field, rotarod)
Water maze (cognition assays requiring visual cues)	Visual impairments, hyperactivity	Confirm intact vision in test subjects. Confirm absence of motor alterations (e.g., swim speed).
Learning and memory requiring food restriction	Low performance due to inadequate restriction (low motivation)	Match subjects for % reduction in body weights due to restriction, relative to pre-restriction
Pre-pulse inhibition of the acoustic startle	Hypoactivity, hearing impairment, seizure activity	Confirm absence of motor differences and intact hearing. Ensure drug or phenotype is not inducing seizure.
Forced swim test, tail suspension test	Hypoactivity, hyperactivity	Increased immobility due to hypoactivity/sedation or reduction in immobility due to hyperactivity
Tail suspension test	Tail climbing, hind limb clasping	Visually confirm no tail climbing and exclude tail climbers
Social behavior	Anxiogenic activity, hyperactivity, hypoactivity	Confirm no alterations in distance traveled. Confirm no alterations in anxiety phenotype.
Nociception (hot plate, Von Frey)	Hyperactivity, hypoactivity, sedation	Increased activity would confound the ability to test for sensitivity to the stimulus. Hypoactivity may reduce the reactivity to the stimulus and confound its interpretation. Confirm lack of hyperactivity in alternative test (e.g, open field, rotarod).
Elevated plus maze, elevated zero maze, light/dark test	Hyperactivity, hypoactivity, sedation, ataxia	Confirm no differences in total entries and/or distance traveled. Confirm lack of ataxia in alternative test (e.g, rotarod).
Novel object recognition	Anxiety, neophobia, hyperactivity, hypoactivity	Ensure no issues of neophobia by pre-assessment of object salience and lack of object bias in independent cohorts which may vary across genotypes. High levels of anxiety and altered motor activity confound the interpretation of this test.
Grip strength	Significant differences in body weight	Force measurements should be normalized to body weight
Rotarod	Hyperactivity, ataxia, body size	Confirm absence of motor differences. Normalize for body weight as a factor.