Neurobiology of Learning and Memory xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Neurobiology of Learning and Memory

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

Rigor and reproducibility in rodent behavioral research

Maria Gulinello[a], Heather A. Mitchell[b], Qiang Chang[b], W. Timothy O'Brien[c], Zhaolan Zhou[c], Ted Abel[c,i], Li Wang[d], Joshua G. Corbin[d], Surabi Veeraragavan[e], Rodney C. Samaco[e], Nick A. Andrews[f], Michela Fagiolini[f], Toby B. Cole[g], Thomas M. Burbacher[g], Jacqueline N. Crawley[h,*]

[a] IDDRC Behavioral Core Facility, Neuroscience Department, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[b] IDD Models Core, Waisman Center, University of Wisconsin Madison, Madison, WI 53705, USA
[c] IDDRC Preclinical Models Core, Children's Hospital of Philadelphia and University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
[d] IDDRC Neurobehavioral Core, Center for Neuroscience Research, Children's National Health System, Washington, DC 20010, USA
[e] IDDRC Neurobehavioral Core, Baylor College of Medicine, Houston, TX 77030, USA
[f] IDDRC Neurodevelopmental Behavior Core, Boston Children's Hospital, Boston, MA 02115, USA
[g] IDDRC Rodent Behavior Laboratory, Center on Human Development and Disability, University of Washington, Seattle, WA 98195, USA
[h] IDDRC Rodent Behavior Core, MIND Institute, University of California Davis School of Medicine, Sacramento, CA 95817, USA
[i] Current affiliation: Iowa Neuroscience Institute, University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

Keywords:
Behavior
Mice
Rats
Behavioral assays
Experimental design
Statistical analysis
Cognitive
Novel object recognition
Best practices
Replication
Rigor
Reproducibility

ABSTRACT

Behavioral neuroscience research incorporates the identical high level of meticulous methodologies and exacting attention to detail as all other scientific disciplines. To achieve maximal rigor and reproducibility of findings, well-trained investigators employ a variety of established best practices. Here we explicate some of the requirements for rigorous experimental design and accurate data analysis in conducting mouse and rat behavioral tests. Novel object recognition is used as an example of a cognitive assay which has been conducted successfully with a range of methods, all based on common principles of appropriate procedures, controls, and statistics. Directors of Rodent Core facilities within Intellectual and Developmental Disabilities Research Centers contribute key aspects of their own novel object recognition protocols, offering insights into essential similarities and less-critical differences. Literature cited in this review article will lead the interested reader to source papers that provide step-by-step protocols which illustrate optimized methods for many standard rodent behavioral assays. Adhering to best practices in behavioral neuroscience will enhance the value of animal models for the multiple goals of understanding biological mechanisms, evaluating consequences of genetic mutations, and discovering efficacious therapeutics.

1. Introduction

Scientists seek the truth. Neuroscientists design our experiments to maximize the chances of discovering fundamental biological principles. Decades of trial and error have yielded ever-improving strategies for effective experimental designs. Unbiased data collection, the correct control groups, sufficient sample sizes, randomized sampling, methodological rigor using gold-standard techniques, maintaining blinded data collection, stringent statistical analyses, and the importance of replicating each finding, are issues taught to graduate students and recommended by journal editors. Grant proposals to the National Institutes of Health in the United States are now required to explicate research strategies that promote unbiased rigor, transparency and reproducibility for generating replicable findings. Rigorous approaches to ensure sufficient statistical power are a priority recently described in the National Institute of Mental Health Request for Information Notice NOT-MH-17-036, (https://grants.nih.gov/grants/guide/notice-files/NOT-MH-17–036.html). Yet many published discoveries are subsequently not reproduced (Collaboration, 2015; Gilmore, Diaz, Wyble, & Yarkoni, 2017; Jarvis & Williams, 2016; Landis & et al., 2012). What are the reasons for the apparent failures of the scientific enterprise to ensure robustness and reproducibility of every discovery?

Principles of high relevance to rodent behavioral research include (a) well-validated assay methods, (b) sufficient sample sizes of randomly selected subject animals, (c) consideration of sex differences as detected by male-female comparisons, (d) consideration of age factors,

* Corresponding author at: MIND Institute and Department of Psychiatry and Behavioral Sciences, University of California Davis School of Medicine, Room 1001A Research II Building 96, 4625 2nd Avenue, Sacramento, CA 95817, USA.
E-mail address: crawley@ucdavis.edu (J.N. Crawley).

https://doi.org/10.1016/j.nlm.2018.01.001
Received 1 November 2017; Received in revised form 22 December 2017; Accepted 3 January 2018

especially for rodent models of neurodevelopmental disorders, along with using age-matched controls; (e) consideration of background strain phenotypes to optimize the choice of parental inbred strain for breeding a new mutant line of mice, and (f) using wildtype littermates as the most appropriate controls for genotype comparisons. Full reporting of the test environment and testing apparatus is key to accurately interpreting results. Employing well-established assays from the extensive behavioral neuroscience literature enables meaningful replication studies across laboratories. Conducting a battery of behavioral assays in a defined sequence that proceeds from the least stressful to the most stressful will further enhance replicability across labs. Beginning with measures of general health can help rule out physical disabilities that impair procedural requirements. Health confounds could introduce artifacts and invalidate the interpretation of results from complex behavioral tests, e.g. a mouse that cannot walk will score poorly on behavioral tests that require locomotion. Conducting two or more assays within the same behavioral domain, e.g. two corroborative social tests or three corroborative learning and memory tasks, may increase the reliability of findings. Approaches to maximally confirm a positive finding include (a) repeating studies with a second independent cohort of mice or rats in the researcher's lab; (b) employing another closely related but non-identical experimental manipulation such as another drug from the same pharmacological class; (c) replicating findings in other labs (Crawley, 2008; Crawley & Paylor, 1997).

In this review, we focus on replicability issues in analyzing rodent behaviors. Behavioral assays have a reputation for high variability, as discussed below. Ideally, highly standardized testing protocols will become widely used. In practice, reaching consensus on methods has proven difficult because of the varieties of available equipment, and varying local conditions. Importantly, the innate behavioral repertoire of mice and rats is influenced by a broad range of environmental factors, including aspects of parenting received from birth through weaning, dominance hierarchies in the home cage, amount of human handling prior to testing, previous testing experiences, olfactory cues from the investigators, and physical properties of the vivarium and laboratory such as lighting, temperature and noises (Crabbe, Wahlsten, & Dudek, 1999; Sorge & et al., 2014; Voelkl & Wurbel, 2016; Wahlsten et al., 2003). Many small but essential details affect the success of each rodent behavioral assay. It is most helpful to learn these essential tips from an expert behavioral neuroscience laboratory, to avoid making common novice mistakes. In fact, innate biological variability is similarly a property of rodent anatomy, physiology, biochemistry, and genetics. Analogous considerations apply to other fields of neuroscience, including imaging, electrophysiology, neurochemistry, optogenetics, inducible pluripotent stem cell phenotypes, and gene expression studies (Gilmore et al., 2017; Marton & Sohal, 2016; Peixoto et al., 2015; Wang, Smith, Murphy, & Cook, 2010; Young-Pearse & Morrow, 2016). Many of the principles presented in this article are applicable across neuroscience research disciplines.

2. Relevance to the goals of the special issue on behavioral analyses of animal models of intellectual and developmental disabilities

In this Special Issue, Directors of Rodent Behavior Cores at NICHD-supported Intellectual and Developmental Disorders Research Centers (IDDRCs) and other behavioral core facilities offer their expertise for conducting rigorous mouse and rat behavioral assays. Behavioral neuroscientists routinely use a variety of standardized rodent behavioral assays, as described in many reviews (Bussey et al., 2012; Crabbe, Phillips, & Belknap, 2010; Crawley, 2007; Cryan & Holmes, 2005; Kazdoba, Leach, & Crawley, 2016a; Kazdoba et al., 2016b; Moser, 2011; Puzzo, Lee, Palmeri, Calabrese, & Arancio, 2014; Wohr & Scattoni, 2013). Rodent behavioral assays fall into four general categories of standardized scoring methods. (1) Fully automated mouse and rat behavioral assays such as open field, acoustic startle, prepulse inhibition, contextual and cued fear conditioning, and operant learning. (2) Semi-automated video tracking assays such as Morris water maze, novel object recognition, and 3-chambered social approach, which are dependent on sensitive parameters, settings, and appropriate statistical interpretations. (3) Observer scored assays such as non-automated elevated plus-maze, forced swim, spontaneous alternation, intradimensional/extradimensional set-shift using olfactory substrates, reciprocal social interactions, and repetitive self-grooming, are potentially subject to unconscious rater bias and require evidence of high inter-rater reliability. (4) Automated assays in which the data sets are very large and complex, such as Intellicage (Krackow & et al., 2010; Robinson & Riedel, 2014) and interesting new machine learning approaches (Hong et al., 2015; Lorbach et al., 2017; Wiltschko et al., 2015), involve massive data acquisition approaches that may introduce ambiguities which limit the interpretations of results. In the present opinion article, we offer suggestions of strategies toward standardizing rodent behavioral assays and ensuring reproducibility. Examples presented focus on novel object recognition, one of the most methodologically variable of the cognitive assays that are commonly used by behavioral neuroscientists to investigate the neurobiology of learning and memory.

2.1. Strategies to ensure reproducibility

The importance of identifying and rigorously assessing functional behavioral outcomes is often underestimated. In many cases, functional and behavioral outcomes are still the primary, and sometime the only, means of diagnosis of intellectual disabilities in humans, as there are many syndromes without unequivocal genetic causes, biomarkers, or pathophysiology. Furthermore, it is fundamentally the amelioration and/or prevention of the negative behavioral sequalae of disease, including such symptoms as pain, depression and cognitive or sensorimotor impairment, that is the true goal of any search for novel therapeutics or underlying mechanisms. Thus, it is essential that functional assays conducted in laboratory animals (as well as in humans) are valid, reliable and reproducible.

In theory, the strategies to ensure reproducibility and reliability in behavioral assays are no different than those necessary to ensure rigor in other fields, and follow the guidelines for good laboratory practice. These include thoroughly researching the field, meticulous record keeping, identifying and minimizing extraneous sources of variability and confounds, limiting experimental error, ensuring validity, having sufficient sample size, optimizing the assay so that neither ceiling nor floor effects limit the assay sensitivity, and having well-trained personnel perform the assay. However, logistically and practically, many factors that have no effect on biochemical assays, for example, time of day of testing or test order effects, can have profound effects on functional and behavioral outcomes in laboratory animals. This is further complicated by the fact that some behavioral domains may be more susceptible to particular types of artifacts than others.

2.2. Example of one widely used cognitive task: novel object recognition

The novel object recognition assay was developed by Ennaceur and colleagues in 1988 (Ennaceur, Cavoy, Costa, & Delacour, 1989; Ennaceur & Delacour, 1988; Ennaceur & Meliani, 1992) in rats and relies on the innate tendency of rodents to preferentially explore novel objects. This is most commonly conducted with one exposure to 2 identical objects in an open field (sample trial, Trial 1, familiarization or training trial) followed by a retention interval and subsequent testing (Trial 2, testing trial, retention trial) in which one of the familiar objects has now been replaced with a novel object. Cognitively intact animals should explore the novel (new) object more than the familiar (old) object. At first glance, the task is deceptively simple and can be implemented without large financial investment in specialized equipment. In practice, there are many variations in the coding parameters,

Table 1
Representative examples of methods used for novel object recognition in mice. **Column 1** lists the NICHD-supported Intellectual and Developmental Disabilities Rodent Core facilities that contributed information about standard methods which have been effective in their studies. **Column 2** confirms that object pairs were pre-tested to ensure that the two objects, presented simultaneously, evoked equal amounts of exploration by control mice. **Column 3** summarizes various protocols used for habituating subject mice to the test arena without objects present. Habituation is designed to reduce the time spent exploring arena surfaces during the novel object test phases. Procedures for habituation vary across facilities for many reasons, including properties of the testing environment and background strain of the subject mice. **Column 4** displays the retention interval, i.e. the length of time between familiarization of the two identical objects and choice between the familiar and the novel object. Retention intervals are usually selected to evaluate aspects of short-term and long-term memory. **Column 5** lists the statistical tests used to determine whether novel object recognition was significant, i.e. more time was spent exploring the novel mouse than the familiar object, for each genotype or for each treatment group. Choice of statistical test varies according to experimental design. **Column 6** provides references for obtaining more detailed information about novel object recognition procedures used by reputable behavioral laboratories. Further descriptions of standard protocols and caveats for conducting novel object recognition assays appear in the text sections under "Strategies to Ensure Reproducibility."

IDDRC site, Rodent Core Manager and/or Director	Were the two objects previously confirmed for equal valence? (Yes/No)	Prior habituation to the empty test chamber (# days, # minutes/day)	Interval between familiarization and recognition sessions (minutes, hours or days)	Data presentation and statistical analyses used	Reference(s)
UC Davis MIND Institute, Crawley	Yes 6 objects in use across studies, object pairs counterbalanced	30 min/day, 1–4 days, varies by strain	1 h or 30 min for short-term memory, 24 h for long-term memory	Repeated Measures ANOVA; One Way ANOVA with Tukey's posthoc for preference and discrimination indices; paired *t*-test for sniff times within-group	Gompers and et al. (2017), Silverman et al. (2013), Stoppel et al. (2017), Yang et al. (2015)
Albert Einstein College of Medicine, Gulinello	Yes, 2 objects used, counterbalanced, validated	Usually 6 min, sometimes no habituation	1–24 h	ANOVA, Chi Square % preference and pass/fail rate	Dai et al. (2010), Vijayanathan et al. (2011)
UW Madison Waisman Center, Mitchell/Chang	Yes, 4 objects used, counterbalanced	Mice pre-handled by experimenter for at least 3 days, habituation to chamber is one 6-min session	3 min	ANOVA	Hullinger et al. (2016)
UPenn, CHOP O'Brien/Abel	Yes	5 min/day, 3–5 days, varies by strain	1 h for short-term memory, 24 h for long-term memory	Two-way ANOVA for gene effect, drug effect, and gene × drug interaction, Tukey post-hoc	Oliveira, Hawk, Abel, and Havekes (2010), Patel and et al. (2014)
Children's National Medical Center, Wang/Corbin	Yes	15 min/day, 4 consecutive days	15 min, 6-hour interval	Two-Way Repeated Measures ANOVA, Newman-Keuls post-hoc	Wang, Jiao, and Dulawa (2011)
Baylor College of Medicine, Veeraragavan/Samaco	Yes	5 min/day,1–4 days	24 h	One-Way ANOVA with LSD or Tukey's posthoc, paired *t*-test	Lu and et al. (2017)
Boston Children's Hospital, Andrews/Fagiolini	Yes	Typically 5 min on day prior to first trial of same objects, sometimes no habituation	10 min, 24 h, depending on mouse model	ANOVA, genotype × object between/within	Arque et al. (2008)
University of Washington, Cole/Burbacher	Yes	Habituation to room > 1 week; pre-handling 3–5 days; habituation to chamber 15–30 min, 1 day prior to testing	1 h, 48 h	One-Way ANOVA or Two-Way ANOVA with Bonferroni's post hoc	Choi et al. (2017), Pan, Chan, Kuo, Storm, and Xia (2012), Wang et al. (2014)

Fig. 1. Examples of object pairs used in novel object recognition testing by IDDRC Rodent Behavior Cores. (A) Left: Orange traffic cone, 7 cm high × 5 cm wide (Amazon.com), and green cylindrical magnet, 7 cm high × 3 cm wide (source: Magneatos, GuideCraft, Amazon.com); photo by Michael Pride, UC Davis MIND Institute Rodent Behavior Core). (B) Premium Big Briks, 3 staggered reclining bricks, ~6 cm high × 3 cm wide (source: Amazon.com # B01N5FGUHB), coral, 5 cm high × 3 cm wide (Safari Ltd, Miami Lake, FL) and treasure chest 2 cm high × 4 cm wide (Safari Ltd, Miami Lake, FL); photo by Melanie Schaffler, UC Davis MIND Institute. Unpublished photos in A and B were contributed by Jacqueline Crawley, UC Davis MIND Institute IDDRC Rodent Behavior Core. (C) Binder clip (source: Office Depot, Washington DC, USA), Open field chamber (source: Accuscan, Columbus, OH, USA); decorated binder clip. Unpublished photos contributed by Li Wang, Children's National Health System (CNHS) Center for Neuroscience IDDRC Neurobehavioral Core. (D) R2D2 toy, plastic Easter egg, gold oval, metal toy car (source: Target), and Lego block (source unknown). Unpublished photos contributed by Heather Mitchell, University of Wisconsin-Madison, Waisman Center, IDD Models Core. (E) Upper left: Plastic shapes (source: Toys“R”Us); photo by Brett Mommer and Zhengui Xia. Upper right: Gloss-painted wooden blocks (source unknown); photo by Christine Cheah and William Catterall. Lower left: small sand-filled and water-filled glass jars (source unknown), photo by Brett Mommer and Zhengui Xia. Lower right: Plastic train whistles and mini color 3x3 cube puzzles (source: Amazon.com), photos provided by Melissa Barker-Haliski and H. Steve White. Assembled photos contributed by Toby Cole, University of Washington IDDRC Mouse Behavior Laboratory. (F) Left: child’s sippy cup, right, baby bottle; middle- ruler for scale. Unpublished photos contributed by Maria Gulinello, Albert Einstein College of Medicine IDDRC Animal Phenotyping Core. (G) Plastic pipe fittings (source: Ace hardware, Porter Square, Cambridge, MA) 7 cm × 2 cm, and glass jar 10 cm × 3.5 cm. Unpublished photo contributed by Nick Andrews, Boston Children’s Hospital IDDRC Neurodevelopmental Behavior Core. (H) LEGO Classic Medium Creative Brick Box and LEGO Duplo Deluxe Box (source: Amazon.com #10,696 and #10,580). Orange, black and yellow tower dimensions: 15 cm tall × 6.5 cm wide × 6.5 cm long. White, red, blue, yellow tower dimensions: 15.5 cm × 10 cm at the base, 7.5 cm wide along the rest of the body. Brown, yellow, green, orange tower dimensions: 16.5 cm tall × 6.5 cm wide × 9.5 cm long. Unpublished photo contributed by Surabi Veeraragavan, Baylor College of Medicine IDDRC Neurobehavioral Core. (I) Bottle, 4.5 cm diam × 16.5 cm tall, clear, filled with white sand and capped (source: Michaels.com). Metal bar, 3.75 cm × 3.75 cm × 15 cm tall, thin aluminum sheet cut and fabricated with glue, filled with white sand (source: in-house fabrications shop). Objects are mounted on a 6.5 cm × 6.5 cm base. Unpublished photos contributed by Tim O’Brien, University of Pennsylvania, Children’s Hospital of Philadelphia IDDRC Neurobehavior Testing Core.

methodological parameters and in calculating cognitive performance (Antunes & Biala, 2012; Bertaina-Anglade, Enjuanes, Morillon, & Drieu la Rochelle, 2006; Bevins & Besheer, 2006; Ennaceur, 2010; Leger et al., 2013). These include whether the animals are habituated to the test arena prior to the assay, how many habituation sessions are conducted and for how long, the shape and nature of the arena, properties of the object pairs, the duration of the familiarization (training, sample) and novel object recognition (test) trials, and the duration of the retention interval. Other critical variables include lighting conditions, handling of the subjects, and nocturnal or diurnal testing.

Table 1 describes methodological parameters that have been used by several IDDRC Rodent Cores in successfully conducting novel object recognition testing of rodent models of human neurodevelopmental disorders with intellectual disabilities. Fig. 1 illustrates object pairs that have proven useful in novel object recognition assays across IDDRC facilities.

Methods for measuring and analyzing performance are also diverse (Antunes & Biala, 2012; Bevins & Besheer, 2006; Leger et al., 2013; Vogel-Ciernia & Wood, 2014). Definitions of exploratory sniffing vary according to scoring system. Automated videotracking systems usually define a zone around the object, e.g. a 2 cm annulus. More sophisticated videotracking systems which detect body shapes further require that the triangular nose shape is pointing in the direction of the object. When scoring manually, sniffing is generally defined as the nose pointing toward the object and within a 1–2 cm distance from the object. As shown in Fig. 2, several outcome measures are in use (Antunes & Biala, 2012; Bevins & Besheer, 2006; Leger et al., 2013; Vogel-Ciernia & Wood, 2014) and include (1) Time spent exploring each object. The absolute time spent exploring the novel object and the familiar object, in seconds, during the recognition test phase provides the most transparent presentation of the raw data. (2) Preference score. The preference score for the recognition test is calculated as (novel object exploration time/(novel object + familiar object exploration time) × 100%. (3) Difference score. The difference score is calculated as time spent exploring the novel object – time spent exploring the familiar object during the recognition test. (4) Discrimination index. The discrimination index is calculated as time spent exploring the novel object – time exploring the familiar)/(total time exploring both novel + familiar). Derived index scores, such as the preference score and discrimination index, correct for individual differences in total exploration. This score also provides an objective value for “failure” as a 50% preference score reflects equal exploration of both novel and familiar objects. However, the effect size can be small, as the average preference score of healthy subjects is typically 60–70% and the mean

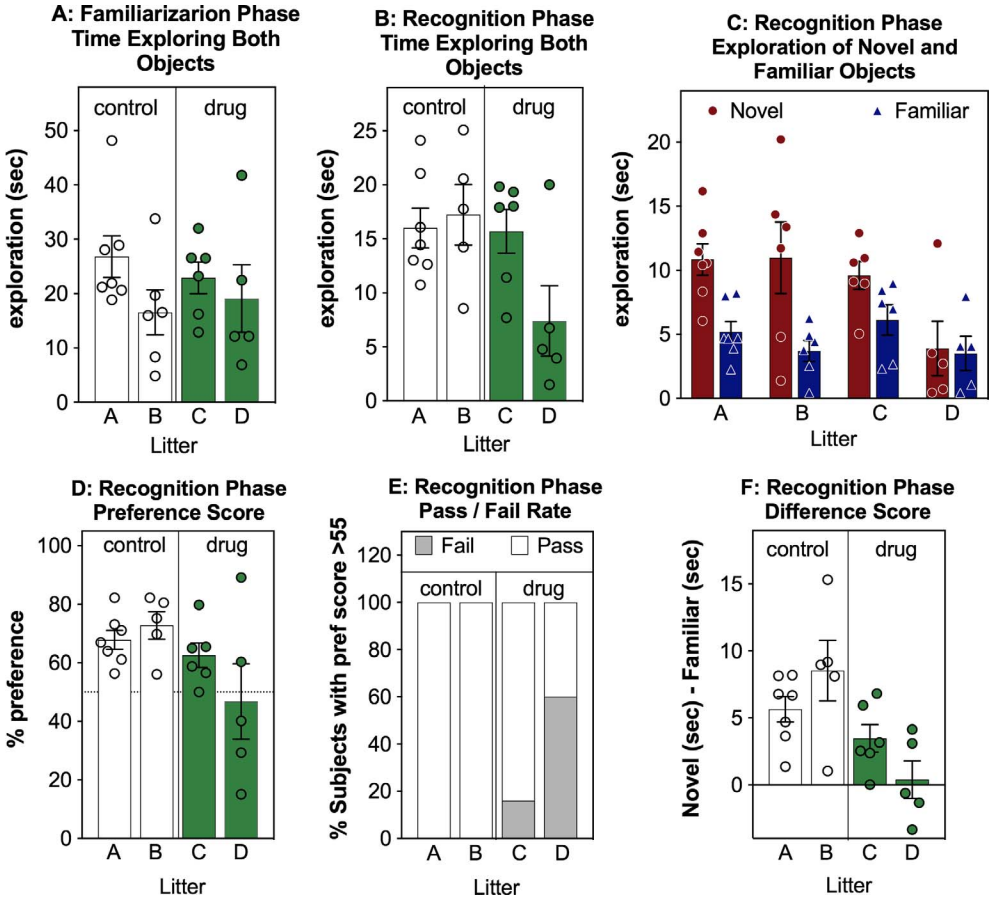

Fig. 2. Common methods for analyzing and illustrating novel object recognition data. (A) Total object exploration of both objects (in seconds) during the familiarization phase (training, Trial 1, sample phase) scored and illustrated separately for each of four litters (X axis- designated A, B, C and D) (B) Total object exploration of both objects (in seconds) during the novel object recognition phase (Test, Trial 2). (C) The absolute exploration of the new (novel) object (red, circle) and the old (familiar) object (blue, triangle), in seconds, during the novel object recognition test phase. (D) Preference score = [(exploration novel object in sec)/(exploration novel + exploration familiar)] × 100 during the novel object recognition test phase. A 50% preference score = the same exploration both the novel (new) and the familiar (old) object). (E) Pass/Fail rate indicates the % of subjects with and without a preference for the novel object, with a preference designated as > 55% preference score during the novel object recognition test phase. (F): Difference score = (exploration of novel object in sec)-(exploration of familiar object in sec) during the novel object recognition test phase. A difference score of 0 = the same exploration both the novel and the familiar object. Data shown are from 4 individual litters of C57BL/6 mice wherein dams were treated during pregnancy with either normal drinking water or a drug in the drinking water. Offspring were tested at 5 mos. Litters A and B are control treated (open bars) and litters C and D are drug treated (green bars). Unpublished data by Gulinello, Einstein, IDDRC core.

score if all animals fail will be about 50%. These derived indices can result in different outcomes and often different variability than absolute measures of exploration (Akkerman, Prickaerts, Steinbusch, & Blokland, 2012b; Akkerman et al., 2012a). (4) Pass/fail rate. Because there is currently no convincing evidence that the absolute magnitude of the preference for the novel object is related to the inherent strength of the memory or extent of cognitive dysfunction, an additional possibility is to treat the data as categorical or binary and illustrate and analyze the percent of subjects passing and failing. “Pass” is calculated as the percentage of subjects exploring the novel object more than the familiar object.

Fig. 2 is an exemplar for illustrative purposes, to demonstrate these several ways of presenting the data, to highlight some critical issues and familiarize the reader with the various advantages and problems associated with each way of presenting and analyzing the data. These data also graphically express a potential source of variability, specifically variability that could be due to litter effects. Discussion of the various methods of analyses is detailed in the statistics section below.

No one single way of illustrating and analyzing the data is superior or intrinsically correct – some methods more accurately illustrate important patterns of the data. Data should be presented in the way that most honestly reflects the true outcome of the experiment and any aspect of the data that would affect interpretation, including showing individual data points in addition to means and error bars. Thus, in Fig. 2D the standard bar chart with error would indicate that only one drug-treated litter tends to perform worse than the control litters, however Fig. 2F, more dramatically illustrates that both litters C and D potentially have lower exploration of the novel object compared to the familiar object. The pass/fail mosaic plot (Fig. 2E) indicates also that a greater proportion of litter D does not have a preference for the novel object than litter C but that both drug treated litters tend to perform worse than control litters. This figure also illustrates the critical need to have sufficient sample size, as in these data there is neither the statistical power to warrant combining litters nor to detect litter effects should they be present.

It is also essential to first establish that the control group exhibits significant novel object recognition, to confirm that the methods are working correctly and that the task parameters are feasible for the specific age, species, and strain (Akkerman et al., 2012a, 2012b). Furthermore, the absolute levels of exploration during the habituation and familiarization phases (Fig. 2A), preceding the novel object recognition phase of the assay (Fig. 2B), are reported as an internal control measure of general exploration, to avoid artifacts and misinterpretations due to motor dysfunction or abnormalities in spontaneous exploration of the environment and objects.

3. Successful novel object recognition methods – What works, what doesn’t, and what factors influence the reproducibility of your cognitive tests

So, did the experiment in Fig. 2 “work”? Did the controls display novel object recognition? Did any patterns appear during the familiarization phase that would indicate exploratory confounds? Was there an effect of the drug? And more importantly, if there was an effect, how likely would it be that the investigator could precisely replicate the drug effects and that others could also replicate these data? The answer in this case is equivocal, as it appears plausible that one litter is more affected by the drug than the other. Sample size in this exemplar is insufficient to make a statistical conclusion one way or another. This example highlights the many ways that an apparently simple task could be difficult to replicate, and some of the reasons why it is so often difficult to replicate published work.

It is clear that there is a range of variables that can affect behavioral and physiological assays in rodents, thus we focus next on the critical

parameters that can affect outcomes in behavioral assays with a focus on the novel object recognition assay. A summary of some of the main factors is listed below, and then discussed in more detail.

- Empirical Factors
  - Training of staff and logistics of measurement
  - Investigator factors
  - Object validations
  - Habituation, cleaning the arena and objects, olfaction
  - Exclusion criteria
  - Retention interval and test duration
- Animal Factors
  - Housing conditions
  - Sex
  - Age
  - Breeding strategy
  - Circadian and seasonal
  - Vendor source
- Experimental Design Factors
  - Blinding
  - Matching and circadian factors
  - Sample sizes

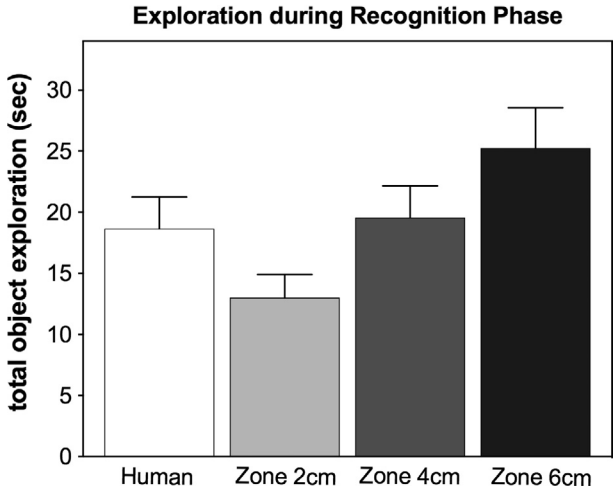

Fig. 4. The size of the zones (X Axis) in automated tracking software can affect apparent object exploration scores. Very small differences of 2 cm in the size of the zone around the object can result in significant differences in recorded exploration ($F_{(2,71)} = 5.4$, $p < .01$). N = 24. Results obtained by a trained human scorer are provided for comparison only. Unpublished data by Gulinello, Einstein, IDDRC core.

3.1. Empirical factors

3.1.1. Accuracy of scoring and reliability of scoring

3.1.1.1. Training and inter-rater reliability. A "novice" experimenter who is well-trained by an expert and has practiced on ≈10 subjects may still have unreliable data compared to an experienced and practiced investigator, resulting in differences in the apparent percentage of subjects passing or failing (Fig. 3). In comparison, trained experimenters have a high degree of concordance when compared to another trained investigator (Fig. 3). This should not be surprising to anyone who has tried to pipette perfect triplicates, insert a good cannula or perform any bench work – there are many ways to make mistakes. Training, expertise and independent validation of scores should be required, as correct and accurate criteria for conducting and scoring behavioral assessments are critical.

3.1.1.2. Manual vs automated scoring. Automated scoring with a tracking system can be a reliable alternative, but still requires extensive training in setting up the automated parameters. For example, a small difference (2 cm) in the size of the zones used to score exploration can result in very large differences in calculated exploration of each novel object (Fig. 4) (Silvers, Harrod, Mactutus, & Booze, 2007). Thus, automated scoring is not necessarily free from experimenter error as it also requires substantial training to achieve valid and reliable scores. It is not really possible to just "set it and forget it." Body size, activity levels and even anxiety levels affect the approach patterns of the subjects and thus the appropriate zone size must be set by a trained experimenter (Rutten et al., 2008) and must furthermore be validated.

Although automated tests are appealing to many researchers and funding agencies, automation is not necessarily more reliable nor more valid, or less dependent on experience on the part of the researcher. Most "automated" assays, including video tracking software, infrared beam detection systems, acoustic startle systems, etc., require proper calibration and understanding of complex software and hardware. Issues include implementing the correct internal controls, substantial datasets requiring extensive manipulation, large numbers of parameters for which small variations drastically change the results, and a knowledge of behavioral psychology principles, that when violated, confound or invalidate the results.

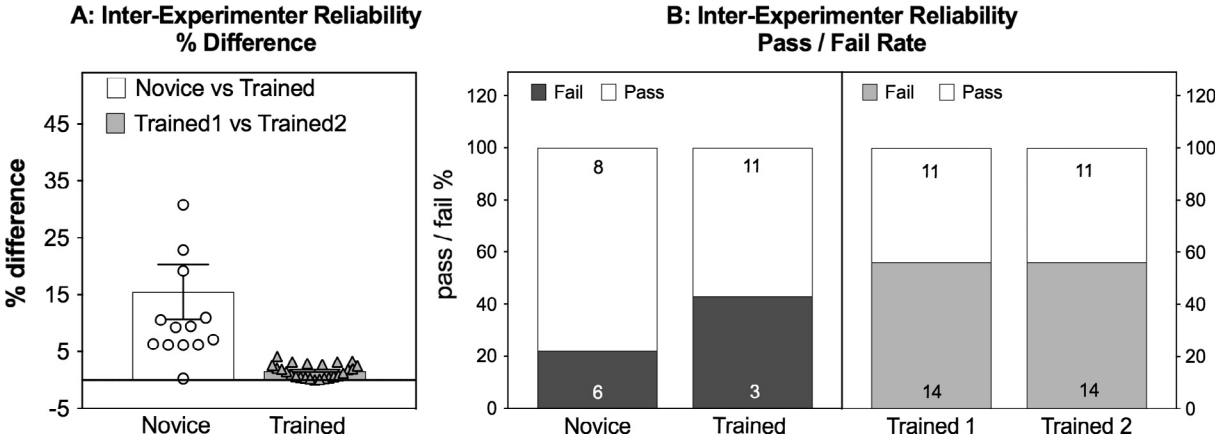

Fig. 3. High degree of inter-experimenter reliability is dependent on sufficient training. (A). The % difference in total exploration in seconds was scored by a novice and a trained experimenter (circles, open bars) when scoring identical movies of the same subjects vs the % difference between 2 trained experimenters also scoring identical movies (triangles, closed bars. % difference = [ABS value (trained-novice)/trained %100] or [ABS value (trained1-trained2)/trained2 %100] (B) The Pass/Fail rate for a novice vs a trained experimenter (left panel) and for 2 trained experimenters (right panel). N for each condition is shown in panel A as individual points in and within the bars in panel B. Note – the trained and novice in panel A and viewed the same movies as did the trained experimenters. The two trained experimenter in panel B viewed identical movies for each comparison, but these experiments were conducted separately, hence the different sample size. Unpublished data by Gulinello, Einstein, IDDRC core.

Another variant is the parameter reported. Some experimenters report the number of entries into the object zones. Others report the duration of time exploring. Although a correlation between the number of entries into the zones and the duration of exploration is expected, there are large deviations from this pattern in sufficient numbers of subjects to potentially invalidate the results if only one parameter is reported.

3.1.1.3. Investigator factors. Many behavioral procedures rely on the test subject responding to novelty in their environment. An often overlooked variable in behavioral testing is the properties of the investigator performing the procedure. To avoid distracting the subject, consideration of the appearance and scent of the investigator is noteworthy. Animal facilities often utilize personal protective equipment as a means of bioprotection of rodent colonies and to limit allergen exposure to investigators. Consistency in the use of scrubs or gowns, gloves, face masks, bonnets, etc. imparts a day-to-day "sameness" to the appearance of the investigator, which may minimize the reaction of the rodent to the investigator. Likewise, changes of perfumes, shampoos, soaps, etc. by an investigator across the days of a procedure might affect behavioral scores in the rodents being tested. Mice appear to respond differently to the sex of investigator, particularly on procedures related to anxiety and pain sensitivity (Sorge et al., 2014), which may be related to odors or handling differences. However, this effect, which has not been demonstrated in the novel object recognition test, lasts for a short time (10–60 min) and can be controlled for by sufficiently habituating the subject to both the testing room and the experimenter (Hanstein et al., 2010). Thus, implementing a standard acclimatization period before testing the subjects is essential.

3.1.2. Object Validations, training period and retention interval

As indicated in Table 1, the most critical factors that affect the outcome of the novel object recognition test, once reliable and valid scoring are established and appropriate experimental design and sample size have been achieved, are (1) validation of the object pairs (no intrinsic preference of either object) and (2) duration of the retention interval (Gulinello, Lebesgue, Jover-Mengual, Zukin, & Etgen, 2006; Ozawa, Yamada, & Ichitani, 2011). (3) Total duration of the test can also be an issue.

3.1.2.1. Object validations. The premise of the novel object recognition assay is that subjects with intact cognitive function will preferentially explore the novel object. Thus, it is assumed that there is no intrinsic preference for either of the objects. The issue is that a highly preferred object will be explored more, regardless of whether it is familiar or novel. Thus, before the test can be conducted reliably, the objects must be validated. If the objects are equally attractive, then animals should explore them both for about the same time during familiarization (training Trial 1, sample trial). The pattern of performance (if subjects pass or fail) will be independent of which object is novel and which is familiar. Further, objects should be counterbalanced during an experiment such that half of each group gets object A as the familiar object and object B as the novel object, and the other half of each group gets B as the familiar object and A as the novel object. Using unvalidated pairs in which a clear preference exists will obscure signficant differences that may otherwise exist (Zhang et al., 2012). This step is also critical to ensure that the two objects are readily distinguishable by the subjects. Behavioral neuroscience laboratories traditionally conduct a series of trial and error validation tests to identify object pairs with sufficient differences and equal valences. It is interesting to note that 3D printing is used by the IDDRC Rodent Core at Washington University at St. Louis to generate highly standardized object pairs (David Wozniak, personal communication). Fig. 1 illustrates some object pairs that have been successfully used by IDDRC Rodent Cores.

The exact stimuli (objects) are not a critical factor in the novel object recognition assay (Busch, Herrmann, Muller, Lenz, & Gruber, 2006). Subjects can be re-tested with new object pairs, as long as all pairs have been validated. High within-subject and between-cohort concordance has been reported for repeated testing of the same subjects with sequential object pairs (Dai et al., 2010; Silverman, Oliver, Karras, Gastrell, & Crawley, 2013; Vijayanathan, Gulinello, Ali, & Cole, 2011; Yang, Lewis, Sarvi, Foley, & Crawley, 2015).

3.1.2.2. Habituation. Various laboratories use different methods for habituation and familiarization in the novel object recognition task. Whether the subjects are habituated to the testing arena prior to the familiarization phase (training, sample phase, Trial 1) and for how long, seems to be idiosyncratic to each successful IDDRC Rodent Core (Table 1). Specific experiments assessing the effect of habituation on subsequent memory performance indicate that prior habituation does not play a great role in affecting cognitive function, nor in general levels of object exploration during the familiarization phase (Leger et al., 2013). However, there may be practical benefits to habituation, particularly in juvenile or highly active subjects. Specifically, habituation may reduce unwanted behaviors such as climbing or leaping that make interpretation and scoring of the assay difficult.

3.1.2.3. Cleaning and the influence of olfaction. Rodents and lagamorphs have exceptional olfactory acuity and use this modality of sensory information to a greater extent that do humans. Mice and rats also scent mark, depositing urinary pheromones in the test arena. Thus, the olfactory environment is potentially an issue in confounding the test. It is worthy of note here that although many assume that mice and rats rely on the visual modality to perform this assay, rodents do not have good visual acuity and do not see color, as they have no fovea and essentially lack cone cells. Olfactory, whisker and tactile modalities provide at least an equal extent of information as does the visual modality. Rodents can reliably perfom novel object preference tests in the dark (Albasser et al., 2010).

Objects that once contained scented contents should generally be avoided as this can alter the preference or avoidance of the object. This information is empirically determined. Many scents, including citrus and perfumes, which are appealing to humans, appear to be aversive to rodents. If it is important to the hypothesis that you do not include regions of the brain that process olfactory information, then it is also critical to establish parameters that prevent the subjects from using olfactory information. Standard practice is to clean the arena with 70% ethanol between trials, and to thoroughly clean the arena and objects after each day's testing. Various dilutions of ethanol, from 10 to 70%, isopropyl alcohol, and antimicrobial cleaning solutions such as vimoba, have been used. It is possible that strongly scented cleaning solutions could influence the novel object recognition results, particularly if insufficient time is allowed for the liquid to evaporate, or the room air ventilation is insufficient to clear odors quickly. However, there is in fact little evidence that the subjects use olfactory cues in the novel object recognition assay (Chan et al., 2017; Hoffman & Basurto, 2013).

3.1.2.4. Criteria for exclusion. For some mutant lines of mice and rats, and in older rodents, exploratory locomotion and total exploration scores may be consistently low. Care should be taken to ensure that there has been adequate exploration of the objects during the familiarization phase (Trial 1, training, sample phase). Preferences based on less than 2–3 sec of data are not technically accurate. Low scores cannot be quantified reproducibly with a stopwatch when scoring manually, are more subject to anomalous entries into a zone when scoring automatically, and do not reflect an adequate sampling of the subject's behavior. Furthermore, when general exploration is low, there is insufficient exploration during habituation for the old objects to become actually familiar, and is thus inadequate for an accurate assessment of a preference for the novel object. There is also a

tendency for those subjects with very low exploration to sit close to a single object. Thus, the behavioral sampling may not reflect an active preference for the proximal object. Data sets can be unreliable if they include many data points with very low preference scores that are generated as the result of inadequate exploration rather than true cognitive deficits. Subjects with very low (< 3–5 sec) exploration "fail" more consistently than subjects with higher exploration. In contrast, in subjects with exploration greater than 3–5 sec during familiarization (Trial 1, training or sample phase), the extent of exploration during the familiarization phase is not correlated with subsequent performance in the novel object recognition phase (Trial 2, test). A requirement for a minimum duration of total object exploration during the familiarization phase, and/or during the novel object recognition test phase, has not been standardized across behavioral neuroscience labs or IDDRCs.

3.1.2.5. Retention interval. The sensitivity of the test, and therefore how likely an effect will be significant and how likely that effect will be reproducible, is also greatly dependent on the retention interval. Generally, longer retention intervals, e.g. 24 h for long-term memory testing, are more "difficult" than shorter retention intervals, e.g. 1 h for short-term memory testing (Gulinello et al., 2006; Leger et al., 2013; Ozawa et al., 2011; Sik, van Nieuwehuyzen, Prickaerts, & Blokland, 2003). The retention interval should be defined and optimized to address the scientific hypothesis about the mutation or treatment. Failing to optimize the retention interval can result in a loss of effect through floor effects (too great a proportion of controls failing) or lack of sensitivity from ceiling effects. Variations in retention intervals between experimenters and different labs can result in an inability to replicate studies through loss of sensitivity.

3.1.2.6. Test duration. The duration of the test is also a parameter which varies between labs and can affect the outcome of the assay. Eventually, the "new" object will also become familiar, and exploration and novel object preference decline over time (Dix & Aggleton, 1999). The optimum time for test duration is 3–5 min in most cases, but can depend on the level of exploration of the subjects, specifically older subjects, subjects that are ill or stressed, etc. Thus, the test duration is another parameter of this apparently simple assay that can greatly affect the performance of the subjects, must be validated empirically, and is often not reported in the methods section.

3.1.3. Animal factors

In addition to variations in the exact methods of conducting and scoring the assay, several other animal factors can affect reliability and reproducibility of the novel object recognition test. These include housing and, vivarium conditions, handling, litter effects, and carryover effects of prior tests or experimental manipulations.

3.1.3.1. Housing conditions. Housing rodents in isolation can have profound effects on the subjects' physiology, health and behavioral sequelae (Chang, Hsiao, Chen, Yu, & Gean, 2015; Chida, Sudo, & Kubo, 2005; Douglas, Varlinskaya, & Spear, 2003; Huang, Liang, Ke, Chang, & Hsieh-Li, 2011; Ibi & et al., 2008; Kwak, Lee, & Kaang, 2009; Lander, Linder-Shacham, & Gaisler-Salomon, 2017; Makinodan & et al., 2016; McLean et al., 2010; Pietropaolo, Singer, Feldon, & Yee, 2008; Pyter et al., 2014; Sakakibara & et al., 2012; Siuda et al., 2014; Talani et al., 2016; Varty, Powell, Lehmann-Masten, Buell, & Geyer, 2006) however see (van Goethem et al., 2012). As shown in Fig. 5, female rats housed in isolation have significantly worse preference scores in the novel object recognition test assessed in a parametric test ($F(1, 20) = 8.7$; $p < .008$) or in a Wilcoxon test ($Z = -2.4 : p < .015$). Chi square (Fischer's exact) analysis of the pass fail/ratio also indicates that a higher proportion of isolated rats fail ($p < .05$). In addition to social housing considerations, husbandry (such as food and water restriction) and exposure to anesthetics, drugs, chronic injections or other stressors, enrichment and surgical manipulations may also alter behavior and apparent performance in the novel object recognition test (Beck & Luine, 1999; Elizalde et al., 2008; Huang, Hayes, & Yang, 2017a; Kawano et al., 2015; Orsini, Buchini, Conversi, & Cabib, 2004; Weiss & Neuringer, 2012; Xiao, Liu, Chen, & Zhang, 2016).

3.1.3.2. Sex differences. Sex differences on novel object recognition have not been routinely seen in control subjects (Fig. 6), however, see (Bettis & Jacobs, 2012; Ghi, Orsetti, Gamalero, & Ferretti, 1999). Females tend to have higher levels of exploration and some groups have found sex differences when absolute level of exploration (in seconds) is used as the measure of cognitive performance as (Bettis & Jacobs, 2012; Ghi et al., 1999). Furthermore, sex differences have been reported in baseline performance and treatment effects on some behavioral assays and not on others. In 1993, the NIH Revitalization Act required the inclusion of women in NIH-funded clinical research. Sex as a biological variable must now be included in NIH grant proposals. Preclinical studies have traditionally not kept up with this policy, as the great majority studies with rodents either use males exclusively, or do not report the sex, or provide no analysis of the similarities and differences when both sexes are used (Zucker & Beery, 2010), despite urging by the NIH to include data from each sex (Clayton & Collins, 2014).

Failure to address sex differences can result in a loss of both reproducibility and translational power. Firstly, single sex studies, or inadequately powered studies which cannot detect sex differences should they occur, result in loss of translation power to half the human population. Secondly, there is the potential for skewed data when both sexes are not equally represented in all treatment groups. Despite the lack of sex differences in the novel object recognition assay, sex differences are evident in many other behavioral domains. Assessing the number of slips in the balance beam, a measure of sensorimotor function and motor coordination, indicates significant differences between males and females in 7 month old C57 mice (Mean number of slips: females = 11 ± 3.1, males = 24 ± 3.0; $F(1, 19) = 9.09$, $p < .007$; unpublished data by Maria Gulinello, Einstein, IDDRC core). Sex difference were found to be significant in the visible platform trials in the Morris water maze (Gulinello et al., 2009) though not in the spatial memory trials (Fritz, Amrein, & Wolfer, 2017; Gulinello et al., 2009). Baseline differences between sexes have been reported for other assays, and mutations, treatment and housing conditions may differentially affect either sex (Chitu et al., 2015; Gulinello, Orman, & Smith, 2003; Hanstein et al., 2010; Huang, Zhou, & Liu, 2017b).

3.1.3.3. Age. Scores for novel object recognition seem to be stable in mice tested between 3 and 4 months of age. Older subjects (> 12 months) tend to explore less and do not tend to perform as well as younger subjects at the same retention interval. When testing older subjects, the duration of the familiarization trial, the duration of the recognition test, and the retention interval must be optimized for the age of the subjects. However, the pattern recognition, or spatial, version of this assay is significantly affected even in 12 month old subjects (Benice, Rizk, Kohama, Pfankuch, & Raber, 2006; Cavoy & Delacour, 1993; Haley, Berteau-Pavy, Berteau-Pavy, & Raber, 2012; Li et al., 2015). Pre-weanling pups and mice younger than 2 months may also have poorer and more unreliable scores on novel object recognition than adults (Anderson et al., 2004). Onset of puberty varies within and between strains, ranging from about 25–43 days in female mice (Pinter, Beda, Csaba, & Gerendai, 2007; Yuan et al., 2012). Concomitant hormone fluctuations and developmental changes can contribute to behavioral differences. Younger mice tend to have generally higher activity levels in a novel environment, more rapidly approach novel objects, and spend more time with a novel object relative to adults (Anderson et al., 2004; Silvers et al., 2007).

3.1.3.4. Breeding strategy, circadian and seasonal effects, vendors. Many additional animal factors require consideration. These include the

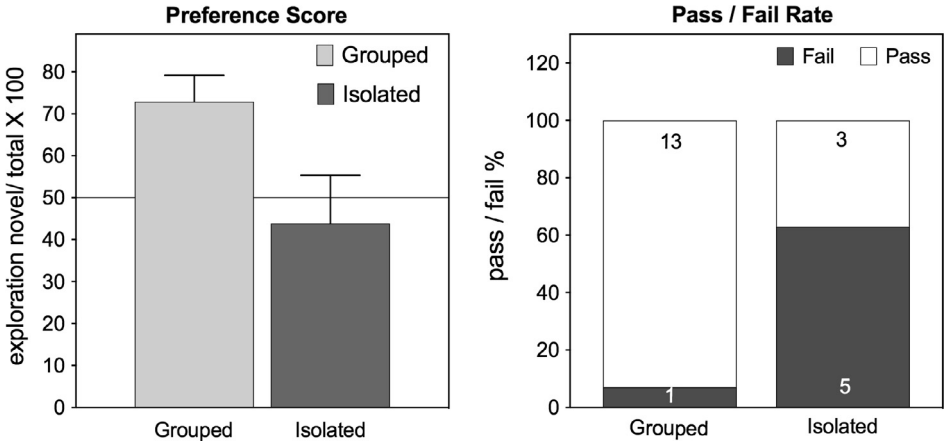

Fig. 5. The effect of isolation on performance in the novel object recognition test (3 min familiarization, 3 min recognition, 30 min retention interval). Female Long-Evans rats were housed in either grouped (2–3 per cage) or isolated conditions for 4–6 weeks and then assessed in the novel object recognition test. Sample sizes are shown in the bars in the pass/fail graph at right. Unpublished data by Gulinello, Einstein, IDDRC core.

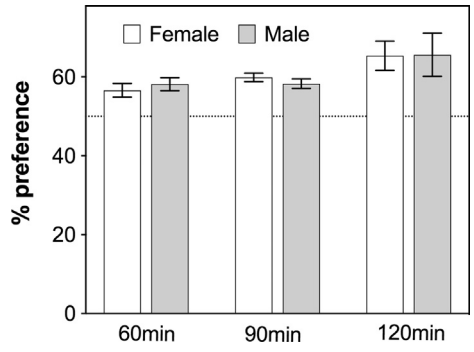

Fig. 6. Absence of sex differences in the novel object recognition test. The graph depicts the performance of male and female mice (total n = 551) on a C57BL/6 background, tested between 3 and 10 months old. Sex differences in performance were not detected at any of the retention intervals. Unpublished data by Gulinello, Einstein, IDDRC core.

number of generations of backcrossing or heterozygote breeding which are conducted for a genetic mutation model, diurnal and circadian effects, and seasonal effects of minor variations in room temperature and humidity during the summer versus winter months. In addition, different commercial vendors supply different substrains of C57BL mice, and may breed mutant lines onto different genetic backgrounds. As mentioned above, the background strain of mice incorporates phenotypes that can directly affect the consequences of a targeted gene mutation.

3.1.4. Experimental design factors

3.1.4.1. Blinding. Unconscious bias in scoring behavioral data is unavoidable. The solution is to ensure that the investigator does not know the genotype of the subject animal, and/or which subject received which treatment. This is usually done by coding the animal genotypes and/or videos with uninformative identification numbers, and/or coding the drug vials with uninformative letters such as A, B, C. Coding is done by another lab member. The rater remains uninformed until after the scoring is completed and the code is broken. Blinding of the experimenter scoring the test is a critical factor as the assessment of behavioral criteria for exploring requires split-second judgments that must be free from unconscious bias. However, in some cases this is logistically difficult to achieve as subjects may have observable phenotypic differences (body weight, fur condition etc.). It is thus useful to have another experimenter independently conduct scoring to confirm scores and observations when the condition of the subjects is grossly visible.

3.1.4.2. Matching and circadian effects. Other important experimental design factors include matching testing blocks. For example, it may take about 8 h to test 20–25 subjects in a typical 2 trial novel object recognition paradigm with a 1 h retention interval. This would mean that some subjects are tested in the beginning of the day and some are tested at the end of the day, introducing potentially confounding circadian effects. Each block of tests should include equivalent numbers of each treatment group, keeping in mind that a treatment group is a single sex, genotype, and/or drug, using Ns that can be tested in about 2 h. Circadian confounds due to time of day of testing or sleep disruption can potentially affect performance in the novel object recognition test (Muller, Fritzsche, & Weinert, 2015; Ruby et al., 2013). Recognition memory may be less susceptible to circadian influences than pattern recognition or object placement memory (Takahashi, Sawa, & Okada, 2013).

3.1.4.3. Sample size. A critical factor in reproducibility and reliability is sample size. Behavioral testing can be time consuming. It is tempting for researchers to "test until significance" – i.e. to keep testing more subjects until an effect is found. This incorrect strategy can yield a small sample size, insufficient for a truly robust data set, or an unreasonably large sample size. In some cases, it is impractical to complete the entire behavioral assay with the sample size as originally designed. It is preferable to combine smaller, manageable subsets to obtain sufficient sample size to define a cohort. Data from the two or three subsets are compared, e.g. scores in the control subgroups should be similar to each other, and scores in the treatment subgroups should be similar to each other, to confirm that the subsets can be combined to form the full cohort with the required N (Gulinello et al., 2009; Hanstein et al., 2010; Yang et al., 2015). In practice, this requires that all animals be subjected to the same exact conditions, i.e. all previous tests in the same order, identical objects, identical test parameters. Good management of lab notes and databases are required, to keep track of all the relevant factors that would prevent substantial intra-cohort variability.

4. Importance of reporting all details in the methods section of publications

Perfunctory methods sections with inadequate details and insufficient citations are a recipe for irreproducible results. There are clearly numerous factors, including specifics of the testing parameters, scoring methods, housing conditions, and age of the subjects, that affect the results obtained and how robust and reliable those results are. Unfortunately, behavioral methods sections are often perfunctory, especially in publications where the primary focus is not primarily behavioral. Suggestions to improve transparency, robustness, and reproducibility of behavioral publications are offered below.

(1) Cite relevant publications from the investigator's lab and others. If the experiment was conducted previously and similar data were

obtained, cite the paper, to let other researchers know that the findings have been replicated.
(2) Methods should be sufficiently detailed to address the major issues affecting the outcomes of the test.
(3) Housing conditions and animal factors should be detailed (barrier or conventional, group housed or isolated).
(4) Citations to a method should focus on papers containing detailed methods and validation. It is best to avoid "ghost" citations that cite a paper that cites a paper that cites a paper that eventually contains the method.
(5) If you haven't read the methods, you haven't read the paper. If you haven't read the paper, please don't cite the paper.
(6) The best advice for investigators and authors whose experience is more molecular or physiological is to confer with behavioral experts before conducting behavioral assays and before publishing the data.
(7) We encourage journal editors to recruit behavioral neuroscientists with appropriate expertise as reviewers of manuscripts in which the behavioral results impact the conclusions.

5. Appropriate statistical tests

Both underpowered studies, from a sample size that is too small, and inappropriately applied statistical tests, can affect the interpretation of the data and thus the likelihood that it will be reproducible. Experiments in live subjects conducted with small sample sizes are prone to artifacts, covariance and confounds that can neither be assessed nor controlled for.

All statistical tests make certain assumptions about the data, that, when violated, can affect the apparent significance. These include assumptions about the distribution and variability and sample size. Nonparametric tests are sometimes called distribution free statistics because they do not require that the data fit a normal distribution. In fact, these are not free of assumptions about the data, and parametric tests offer many advantages. Within their assumptions, parametric tests are robust and have greater power efficiency, e.g. relative to sample size there is higher statistical power. Parametric tests are also more flexible and can provide unique data, specifically about interactions between factors and interactions of the factors over time. Parametric tests are also arguably relatively resistant to violations of the assumption of normality (due to the Central Limit Theorem) provided the sample size is adequate – such as N = 20–30 (Kwak & Kim, 2017).

Here we summarize some general guidelines and strategies for obtaining the highest quality and most rigorous behavioral data.

- The importance of core facilities and expert advice in conducting these assays.

Very few researchers would embark on creating a new mutant mouse or conducting HPLC analysis of samples without appropriate training, literature searches, pilot studies, optimization, appropriate validation studies, internal and external controls and consultation with experts. In contrast, experimenters sometimes think that behavioral assays do not require a high level of expertise. Core facilities and research centers are designed to prevent "re-inventing the wheel", with all the attendant pitfalls that this entails. Core facilities are uniquely able to validate existing methods and equipment and maintain databases over long periods of time, which are conducive to analyses of replications, and can examine meta-data in a manner usually inaccessible to a single researcher or lab. Capabilities of core facilities include assessment of internal controls such as locomotor activity confounds which may impact total object exploration, and retention intervals that set the level of challenge for cognitive tests. Behavioral neuroscience experts routinely evaluate procedural control measures in a new line of mice, since performance on the procedures of a cognitive task can directly affect scores.

In contrast to analytical chemistry methods in which internal and external standards are included within the assay, internal and external standards for behavioral research generally rely on consistency of data from control groups across time, using the maintained databases and meta-data analyses. For the novel object recognition test, internal controls may include (a) assessment of performance with no retention interval, to distinguish non-specific sensorimotor deficits from cognitive dysfunction (Gulinello et al., 2006; Li, Vijayanathan, Gulinello, & Cole, 2010); (b) analysis of counterbalanced object data, to ensure that no object bias exists, even after objects have been previously validated; (c) analysis of exploration levels during the familiarization trial, (d) inclusion of more than one retention interval (Gulinello et al., 2006), and (e) assessment of general locomotor activity, rearing, grooming etc. External controls include (a) evaluation of the consistency of data from control groups across time, (b) between-cohort analyses, and (c) comparison to other assays in analogous behavioral domains (Gulinello et al., 2009) . Does the subject have cognitive impairment, depression-like behavior or tactile insensitivity? Or is that subject simply moving less? One single test can seldom answer that question. Molecular biologists, physiologists and biochemists use internal controls and multiple methods, which each have limitations. Behavioral investigators can and should similarly differentiate between procedural performance and cognitive abilities, and detect likely confounds.

- Honest illustration of the data set

Just as scientists have been admonished for less than honest Western blots and immunohistochemistry images (Neill, 2006), so must behavioral neuroscientists strive to illustrate and analyze their data with scrupulous integrity. In the words of physicist Richard Feynman, "We've learned from experience that the truth will come out. Other experimenters will repeat your experiment and find out whether you were wrong or right. Nature's phenomena will agree or they'll disagree with your theory. And, although you may gain some temporary fame and excitement, you will not gain a good reputation as a scientist if you haven't tried to be very careful in this kind of work. And it's this type of integrity, this kind of care not to fool yourself, that is missing to a large extent" (Feynman & Leighton, 1985).

- Empirical data to test the critical parameters

To what extent should rodents be habituated to the arena/test environment? Should the test session have a duration of 5 min or 20 min? Numerous examples of methodological variations appear in the literature, and should thus use appropriate caution about setting up a new experiment based on one paper. These are empirical questions, with empirical answers. While JOVE and various protocol journals can provide a "quick start guide" that can be invaluable to setting up an unfamiliar paradigm, there is no substitute for testing it, researching it and validating it.

6. Conclusions

Ultimately the success of a rigorous experimental design will be judged by the replicability of its findings. Especially when the rodent behavioral phenotype is applied as a translational tool to evaluate potential clinical treatments, well-replicated and highly robust phenotypes are necessary to detect drug responses, over and above innate biological variability (Begley & Ellis, 2012; Cole et al., 2013; Drucker, 2016; Schulz, Cookson, & Hausmann, 2016; Silverman et al., 2012). One reasonable progression of replications to confirm the universal strength of a finding is: (1) Replication within a lab, repeating the same procedures in two independent cohorts of animals; (2) Evaluations of a range of related behavioral assays, e.g. four learning and memory tasks, or three sociability tests; (3) Replication across labs, each repeating approximately the same experiments in mice or rats with the same

mutation or treatment. In cases where results replicate well within one laboratory but not universally, it is reasonable to assume that methods specific to one lab will require modifications in other labs, to validate the assay. It is useful for labs to look carefully at the experimental parameters and conditions originally reported, and even to take the time to contact the authors, to understand if there are any crucial differences that may determine whether an effect is consistently detected; (4) Comparisons across mutant lines, e.g. using mice with different mutations in the same gene, or mutations in functionally related genes, or a different drug in the same pharmacological class; (5) Comparisons across species, e.g. mice, rats, and non-human primates. Findings that replicate across these stringent criteria would provide high confidence that the animal data are strong enough to inform consideration of a clinical trial.

Issues surrounding the translational value of preclinical animal models as predictive translational tools for clinical trials have been extensively discussed (Belin, Belin-Rauscent, Everitt, & Dalley, 2016; Flier, 2017; Geyer, 2008; Jablonski, Schreiber, Westbrook, Brennan, & Stanton, 2013; Kas & et al., 2014; Kazdoba et al., 2016b; Lynch, Palmer, & Gall, 2011; McGraw, Ward, & Samaco, 2017; Robbins, 2017; Sarter, 2006; Schulz et al., 2016; Snyder & et al., 2016; Spooren, Lindemann, Ghosh, & Santarelli, 2012). Although non-predictive animal studies are one factor, clinical trials fail for many other reasons. These include toxicity and other human safety concerns, dose-response pharmacokinetic variability across human subjects, brain bioavailability of the drug, inappropriate aspects of study design such as age of subjects at treatment onset, characteristics selected for patient stratification such as IQ and language abilities, and choice of primary outcome measures (Insel & et al., 2013; Kola & Landis, 2004; Lythgoe et al., 2016).

In addition, rodents and humans may have innate differences in critical pharmacodynamic, pharmacokinetic, metabolic, immune, and lifespan parameters. Sanchez, Asin, & Artigas, 2015 provides a comprehensive example of pharmacokinetic and pharmacodynamic differences in drug action between rodents and humans. Vortioxetine, a novel, multimodal antidepressant, displays a binding affinity profile to specific serotonin receptors which differs considerably between rodents and humans, as does its absolute oral availability and elimination half-life (Sanchez et al., 2015).

“It is better (and cheaper) that potential targets be discarded at the preclinical level should they prove ineffective.” (Perry & Lawrence, 2017). In terms of financial investment in drug development, conducting rigorous animal studies may be cost-effective. The highest quality of preclinical rodent data will maximize predictive validity. Conversely, findings of no replicable preclinical efficacy may be sufficient to disprove the hypothesized target mechanism, thereby saving future expenditures. However, from the point of view of academic researchers dependent on limited grant funding, well-designed replication studies are expensive, time consuming, and difficult to support within current NIH grant budgets.

Of course, we recognize the conundrum. Research is a costly, time-intensive endeavor. On the other hand, public and private support for financing scientific research depends on confidence that results are trustworthy. In principle, the scientific method is self-correcting. Successful scientists maintain a high level of motivation for the long, hard slog to generate important, pristine findings. Enthusiasm can wane when consistently negative findings are obtained, which cannot be published in good journals. Particularly discouraging is the common scenario wherein careful researchers are scooped by competing labs who publish less rigorous data. A major issue is the difficulty of publishing negative data, particularly failures to replicate, especially when the initial paper is from a prominent laboratory and appears in a high-profile journal (Button & Munafo, 2014; Macleod & et al., 2015).

It is heartening to see the reproducibility issue at the forefront of debate in journals and at major funding agencies (Baker, 2017; Landis et al., 2012; McNutt, 2014a, 2014b). Approaches to improve rigor, transparency and reproducibility of data are now in place in many venues (Collins & Tabak, 2014; Lithgow, Driscoll, & Phillips, 2017; McNutt, 2014a, 2014b; Moher, Simera, Schulz, Hoey, & Altman, 2008). Researchers can certainly be incentivized to invest their limited funding in conducting rigorous experimental designs and replication studies. Labs will be motivated to repeat positive findings in a second experiment before considering publication, when their funding agency supports the replication study, and when editors of high profile journals prioritize manuscripts that incorporate replication studies. Such strategies may ameliorate the “reproducibility crisis” to a great extent. Promoting reproducibility goals could go a long way towards maximizing our discoveries of biological truths.

Acknowledgments

We thank Dr. Tracy King, NICHD, for her leadership and encouragement of the cross-IDDRC comparisons described in Table 1 and Fig. 1. Supported by U54HD090260 (MG); U54 HD090256 (QC); U54HD086984 (WTO, ZZ, TA); U54HD090257 (LW, JGC); U54HD083092 (SV, RCS); U54HD090255 (NAA, MF); U54HD083091 (TBC, TMB); U54HD079125 (JNC).

References

Akkerman, S., Blokland, A., Reneerkens, O., van Goethem, N. P., Bollen, E., Gijselaers, H. J., et al. (2012a). Object recognition testing: Methodological considerations on exploration and discrimination measures. *Behavioural Brain Research, 232*, 335–347.

Akkerman, S., Prickaerts, J., Steinbusch, H. W., & Blokland, A. (2012b). Object recognition testing: Statistical considerations. *Behavioural Brain Research, 232*, 317–322.

Albasser, M. M., Chapman, R. J., Amin, E., Iordanova, M. D., Vann, S. D., & Aggleton, J. P. (2010). New behavioral protocols to extend our knowledge of rodent object recognition memory. *Learning & Memory, 17*, 407–419.

Anderson, M. J., Barnes, G. W., Briggs, J. F., Ashton, K. M., Moody, E. W., Joynes, R. L., et al. (2004). Effects of ontogeny on performance of rats in a novel object-recognition task. *Psychological Reports, 94*, 437–443.

Antunes, M., & Biala, G. (2012). The novel object recognition memory: Neurobiology, test procedure, and its modifications. *Cognitive Processing, 13*, 93–110.

Arque, G., Fotaki, V., Fernandez, D., Martinez de Lagran, M., Arbones, M. L., & Dierssen, M. (2008). Impaired spatial learning strategies and novel object recognition in mice haploinsufficient for the dual specificity tyrosine-regulated kinase-1A (Dyrk1A). *PLoS ONE, 3*, e2575.

Baker, M. (2017). Reproducibility: Check your chemistry. *Nature, 548*, 485–488.

Beck, K. D., & Luine, V. N. (1999). Food deprivation modulates chronic stress effects on object recognition in male rats: Role of monoamines and amino acids. *Brain Research, 830*, 56–71.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Belin, D., Belin-Rauscent, A., Everitt, B. J., & Dalley, J. W. (2016). In search of predictive endophenotypes in addiction: Insights from preclinical research. *Genes, Brain and Behavior, 15*, 74–88.

Benice, T. S., Rizk, A., Kohama, S., Pfankuch, T., & Raber, J. (2006). Sex-differences in age-related cognitive decline in C57BL/6J mice associated with increased brain microtubule-associated protein 2 and synaptophysin immunoreactivity. *Neuroscience, 137*, 413–423.

Bertaina-Anglade, V., Enjuanes, E., Morillon, D., & Drieu la Rochelle, C. (2006). The object recognition task in rats and mice: A simple and rapid model in safety pharmacology to detect amnesic properties of a new chemical entity. *Journal of Pharmacological and Toxicological Methods, 54*, 99–105.

Bettis, T., & Jacobs, L. F. (2012). Sex differences in object recognition are modulated by object similarity. *Behavioural Brain Research, 233*, 288–292.

Bevins, R. A., & Besheer, J. (2006). Object recognition in rats and mice: A one-trial non-matching-to-sample learning task to study 'recognition memory'. *Nature Protocols, 1*, 1306–1311.

Busch, N. A., Herrmann, C. S., Muller, M. M., Lenz, D., & Gruber, T. (2006). A cross-laboratory study of event-related gamma activity in a standard object recognition paradigm. *Neuroimage, 33*, 1169–1177.

Bussey, T. J., Holmes, A., Lyon, L., Mar, A. C., McAllister, K. A., Nithianantharajah, J., et al. (2012). New translational assays for preclinical modelling of cognition in schizophrenia: The touchscreen testing method for mice and rats. *Neuropharmacology, 62*, 1191–1203.

Button, K. S., & Munafo, M. R. (2014). Incentivising reproducible research. *Cortex, 51*, 107–108.

Cavoy, A., & Delacour, J. (1993). Spatial but not object recognition is impaired by aging in rats. *Physiology & Behavior, 53*, 527–530.

Chan, W., Singh, S., Keshav, T., Dewan, R., Eberly, C., Maurer, R., et al. (2017). Mice lacking M1 and M3 muscarinic acetylcholine receptors have impaired odor discrimination and learning. *Frontiers in Synaptic Neuroscience, 9*, 4.

Chang, C. H., Hsiao, Y. H., Chen, Y. W., Yu, Y. J., & Gean, P. W. (2015). Social isolation-induced increase in NMDA receptors in the hippocampus exacerbates emotional

dysregulation in mice. *Hippocampus, 25*, 474–485.
Chida, Y., Sudo, N., & Kubo, C. (2005). Social isolation stress exacerbates autoimmune disease in MRL/lpr mice. *Journal of Neuroimmunology, 158*, 138–144.
Chitu, V., Gokhan, S., Gulinello, M., Branch, C. A., Patil, M., Basu, R., et al. (2015). Phenotypic characterization of a Csf1r haploinsufficient mouse model of adult-onset leukodystrophy with axonal spheroids and pigmented glia (ALSP). *Neurobiology of Diseases, 74*, 219–228.
Choi, W. S., Kim, H. W., Tronche, F., Palmiter, R. D., Storm, D. R., & Xia, Z. (2017). Conditional deletion of Ndufs4 in dopaminergic neurons promotes Parkinson's disease-like non-motor symptoms without loss of dopamine neurons. *Scientific Reports, 7*, 44989.
Clayton, J. A., & Collins, F. S. (2014). NIH to balance sex in cell and animal studies. *Nature, 509*, 282–283.
Cole, P. D., Vijayanathan, V., Ali, N. F., Wagshul, M. E., Tanenbaum, E. J., Price, J., et al. (2013). Memantine protects rats treated with intrathecal methotrexate from developing spatial memory deficits. *Clinical Cancer Research, 19*, 4446–4454.
Collaboration, O. S. (2015). Psychology. Estimating the reproducibility of psychological science. *Science* (pp. 349) aac4716.
Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505*, 612–613.
Crabbe, J. C., Phillips, T. J., & Belknap, J. K. (2010). The complexity of alcohol drinking: Studies in rodent genetic models. *Behavior Genetics, 40*, 737–750.
Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science, 284*, 1670–1672.
Crawley, J. N. (2007). *What's wrong with my mouse? Behavioral phenotyping of transgenic and knockout mice.* Hoboken, New Jersey: John Wiley & Sons Inc.
Crawley, J. N. (2008). Behavioral phenotyping strategies for mutant mice. *Neuron, 57*, 809–818.
Crawley, J. N., & Paylor, R. (1997). A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice. *Hormones and Behavior, 31*, 197–211.
Cryan, J. F., & Holmes, A. (2005). The ascent of mouse: Advances in modelling human depression and anxiety. *Nature Reviews. Drug Discovery, 4*, 775–790.
Dai, M., Reznik, S. E., Spray, D. C., Weiss, L. M., Tanowitz, H. B., Gulinello, M., et al. (2010). Persistent cognitive and motor deficits after successful antimalarial treatment in murine cerebral malaria. *Microbes and Infection, 12*, 1198–1207.
Dix, S. L., & Aggleton, J. P. (1999). Extending the spontaneous preference test of recognition: Evidence of object-location and object-context recognition. *Behavioural Brain Research, 99*, 191–200.
Douglas, L. A., Varlinskaya, E. I., & Spear, L. P. (2003). Novel-object place conditioning in adolescent and adult male and female rats: Effects of social isolation. *Physiology & Behavior, 80*, 317–325.
Drucker, D. J. (2016). Never waste a good crisis: Confronting reproducibility in translational research. *Cell Metabolism, 24*, 348–360.
Elizalde, N., Gil-Bea, F. J., Ramirez, M. J., Aisa, B., Lasheras, B., Del Rio, J., et al. (2008). Long-lasting behavioral effects and recognition memory deficit induced by chronic mild stress in mice: Effect of antidepressant treatment. *Psychopharmacology (Berl), 199*, 1–14.
Ennaceur, A. (2010). One-trial object recognition in rats and mice: Methodological and theoretical issues. *Behavioural Brain Research, 215*, 244–254.
Ennaceur, A., Cavoy, A., Costa, J. C., & Delacour, J. (1989). A new one-trial test for neurobiological studies of memory in rats. II: Effects of piracetam and pramiracetam. *Behavioural Brain Research, 33*, 197–207.
Ennaceur, A., & Delacour, J. (1988). A new one-trial test for neurobiological studies of memory in rats. 1: Behavioral data. *Behavioural Brain Research, 31*, 47–59.
Ennaceur, A., & Meliani, K. (1992). A new one-trial test for neurobiological studies of memory in rats. III. Spatial vs. non-spatial working memory. *Behavioural Brain Research, 51*, 83–92.
Feynman, R., & Leighton, R. (1985). *Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character*. USA: W.W. Norton.
Flier, J. S. (2017). Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy. *Molecular Metabolism, 6*, 2–9.
Fritz, A. K., Amrein, I., & Wolfer, D. P. (2017). Similar reliability and equivalent performance of female and male mice in the open field and water-maze place navigation task. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 175*, 380–391.
Geyer, M. A. (2008). Developing translational animal models for symptoms of schizophrenia or bipolar mania. *Neurotoxicity Research, 14*, 71–78.
Ghi, P., Orsetti, M., Gamalero, S. R., & Ferretti, C. (1999). Sex differences in memory performance in the object recognition test. Possible role of histamine receptors. *Pharmacology, Biochemistry and Behavior, 64*, 761–766.
Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences, 1396*, 5–18.
Gompers, A. L., et al. (2017). Germline Chd8 haploinsufficiency alters brain development in mouse. *Nature Neuroscience, 20*, 1062–1073.
Gulinello, M., Gertner, M., Mendoza, G., Schoenfeld, B. P., Oddo, S., LaFerla, F., et al. (2009). Validation of a 2-day water maze protocol in mice. *Behavioural Brain Research, 196*, 220–227.
Gulinello, M., Lebesgue, D., Jover-Mengual, T., Zukin, R. S., & Etgen, A. M. (2006). Acute and chronic estradiol treatments reduce memory deficits induced by transient global ischemia in female rats. *Hormones and Behavior, 49*, 246–260.
Gulinello, M., Orman, R., & Smith, S. S. (2003). Sex differences in anxiety, sensorimotor gating and expression of the alpha4 subunit of the GABAA receptor in the amygdala after progesterone withdrawal. *European Journal of Neuroscience, 17*, 641–648.
Haley, G. E., Berteau-Pavy, F., Berteau-Pavy, D., & Raber, J. (2012). Novel image-novel location object recognition task sensitive to age-related cognitive decline in non-demented elderly. *Age (Dordr), 34*, 1–10.
Hanstein, R., Zhao, J. B., Basak, R., Smith, D. N., Zuckerman, Y. Y., Hanani, M., et al. (2010). Focal inflammation causes carbenoxolone-sensitive tactile hypersensitivity in mice. *The Open Pain Journal, 3*, 123–133.
Hoffman, K. L., & Basurto, E. (2013). One-trial object recognition memory in the domestic rabbit (Oryctolagus cuniculus) is disrupted by NMDA receptor antagonists. *Behavioural Brain Research, 250*, 62–73.
Hong, W., Kennedy, A., Burgos-Artizzu, X. P., Zelikowsky, M., Navonne, S. G., Perona, P., et al. (2015). Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proceedings of the National Academy of Sciences of United States, 112*, E5351–5360.
Huang, L., Hayes, S., & Yang, G. (2017). Long-lasting behavioral effects in neonatal mice with multiple exposures to ketamine-xylazine anesthesia. *Neurotoxicology and Teratology, 60*, 75–81.
Huang, H. J., Liang, K. C., Ke, H. C., Chang, Y. Y., & Hsieh-Li, H. M. (2011). Long-term social isolation exacerbates the impairment of spatial working memory in APP/PS1 transgenic mice. *Brain Research, 1371*, 150–160.
Huang, Q., Zhou, Y., & Liu, L. Y. (Zhou et al., 2017b). Effect of post-weaning isolation on anxiety- and depressive-like behaviors of C57BL/6J mice. *Experimental Brain Research, 235*, 2893–2899.
Hullinger, R., Li, M., Wang, J., Peng, Y., Dowell, J. A., Bomba-Warczak, E., et al. (2016). Increased expression of AT-1/SLC33A1 causes an autistic-like phenotype in mice by affecting dendritic branching and spine formation. *Journal of Experimental Medicine, 213*, 1267–1284.
Ibi, D., et al. (2008). Social isolation rearing-induced impairment of the hippocampal neurogenesis is associated with deficits in spatial memory and emotion-related behaviors in juvenile mice. *Journal of Neurochemistry, 105*, 921–932.
Insel, T. R., et al. (2013). Innovative solutions to novel drug development in mental health. *Neuroscience & Biobehavioral Reviews, 37*, 2438–2444.
Jablonski, S. A., Schreiber, W. B., Westbrook, S. R., Brennan, L. E., & Stanton, M. E. (2013). Determinants of novel object and location recognition during development. *Behavioural Brain Research, 256*, 140–150.
Jarvis, M. F., & Williams, M. (2016). Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences, 37*, 290–302.
Kas, M. J., et al. (2014). Assessing behavioural and cognitive domains of autism spectrum disorders in rodents: Current status and future perspectives. *Psychopharmacology (Berl), 231*, 1125–1146.
Kawano, T., Eguchi, S., Iwata, H., Tamura, T., Kumagai, N., & Yokoyama, M. (2015). Impact of preoperative environmental enrichment on prevention of development of cognitive impairment following abdominal surgery in a rat model. *Anesthesiology, 123*, 160–170.
Kazdoba, T. M., Leach, P. T., & Crawley, J. N. (2016a). Behavioral phenotypes of genetic mouse models of autism. *Genes, Brain and Behavior, 15*, 7–26.
Kazdoba, T. M., Leach, P. T., Yang, M., Silverman, J. L., Solomon, M., & Crawley, J. N. (2016b). Translational mouse models of autism: advancing toward pharmacological therapeutics. *Current Topics in Behavioral Neurosciences, 28*, 1–52.
Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery, 3*, 711–715.
Krackow, S., et al. (2010). Consistent behavioral phenotype differences between inbred mouse strains in the IntelliCage. *Genes, Brain and Behavior, 9*, 722–731.
Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology, 70*, 144–156.
Kwak, C., Lee, S. H., & Kaang, B. K. (2009). Social isolation selectively increases anxiety in mice without affecting depression-like behavior. *Korean Journal of Physiology & Pharmacology, 13*, 357–360.
Lander, S. S., Linder-Shacham, D., & Gaisler-Salomon, I. (2017). Differential effects of social isolation in adolescent and adult mice on behavior and cortical gene expression. *Behavioural Brain Research, 316*, 245–254.
Landis, S. C., et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature, 490*, 187–191.
Leger, M., Quiedeville, A., Bouet, V., Haelewyn, B., Boulouard, M., Schumann-Bard, P., et al. (2013). Object recognition test in mice. *Nature Protocols, 8*, 2531–2537.
Li, Y., Abdourahman, A., Tamm, J. A., Pehrson, A. L., Sanchez, C., & Gulinello, M. (2015). Reversal of age-associated cognitive deficits is accompanied by increased plasticity-related gene expression after chronic antidepressant administration in middle-aged mice. *Pharmacology, Biochemistry and Behavior, 135*, 70–82.
Li, Y., Vijayanathan, V., Gulinello, M., & Cole, P. D. (2010). Intrathecal methotrexate induces focal cognitive deficits and increases cerebrospinal fluid homocysteine. *Pharmacology, Biochemistry and Behavior, 95*, 428–433.
Lithgow, G. J., Driscoll, M., & Phillips, P. (2017). A long journey to reproducible results. *Nature, 548*, 387–388.
Lorbach, M., Kyriakou, E. I., Poppe, R., van Dam, E. A., Noldus, L., & Veltkamp, R. C. (2017). Learning to recognize rat social behavior: Novel dataset and cross-dataset application. *Journal of Neuroscience Methods*.
Lu, H. C., et al. (2017). Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. *Nature Genetics, 49*, 527–536.
Lynch, G., Palmer, L. C., & Gall, C. M. (2011). The likelihood of cognitive enhancement. *Pharmacology, Biochemistry and Behavior, 99*, 116–129.
Lythgoe, M. P., Rhodes, C. J., Ghataorhe, P., Attard, M., Wharton, J., & Wilkins, M. R. (2016). Why drugs fail in clinical trials in pulmonary arterial hypertension, and strategies to succeed in the future. *Pharmacology & Therapeutics, 164*, 195–203.
Macleod, M. R., et al. (2015). Risk of bias in reports of in vivo research: A focus for improvement. *PLoS Biology, 13*, e1002273.
Makinodan, M., et al. (2016). Social isolation impairs remyelination in mice through

modulation of IL-6. *FASEB Journal, 30*, 4267–4274.
Marton, T. F., & Sohal, V. S. (2016). Of mice, men, and microbial opsins: How optogenetics can help hone mouse models of mental illness. *Biological Psychiatry, 79*, 47–52.
McGraw, C. M., Ward, C. S., & Samaco, R. C. (2017). Genetic rodent models of brain disorders: Perspectives on experimental approaches and therapeutic strategies. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 175*, 368–379.
McLean, S., Grayson, B., Harris, M., Protheroe, C., Woolley, M., & Neill, J. (2010). Isolation rearing impairs novel object recognition and attentional set shifting performance in female rats. *Journal of Psychopharmacology, 24*, 57–63.
McNutt, M. (2014a). Journals unite for reproducibility. *Science, 346*, 679.
McNutt, M. (2014b). Reproducibility. *Science, 343*, 229.
Moher, D., Simera, I., Schulz, K. F., Hoey, J., & Altman, D. G. (2008). Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC Medicine, 6*, 13.
Moser, V. C. (2011). Functional assays for neurotoxicity testing. *Toxicologic Pathology, 39*, 36–45.
Muller, L., Fritzsche, P., & Weinert, D. (2015). Novel object recognition of Djungarian hamsters depends on circadian time and rhythmic phenotype. *Chronobiology International, 32*, 458–467.
Neill, U. S. (2006). Stop misbehaving!. *Journal of Clinical Investigation, 116*, 1740–1741.
Oliveira, A. M., Hawk, J. D., Abel, T., & Havekes, R. (2010). Post-training reversible inactivation of the hippocampus enhances novel object recognition memory. *Learning & Memory, 17*, 155–160.
Orsini, C., Buchini, F., Conversi, D., & Cabib, S. (2004). Selective improvement of strain-dependent performances of cognitive tasks by food restriction. *Neurobiology of Learning and Memory, 81*, 96–99.
Ozawa, T., Yamada, K., & Ichitani, Y. (2011). Long-term object location memory in rats: Effects of sample phase and delay length in spontaneous place recognition test. *Neuroscience Letters, 497*, 37–41.
Pan, Y. W., Chan, G. C., Kuo, C. T., Storm, D. R., & Xia, Z. (2012). Inhibition of adult neurogenesis by inducible and targeted deletion of ERK5 mitogen-activated protein kinase specifically in adult neurogenic regions impairs contextual fear extinction and remote fear memory. *Journal of Neuroscience, 32*, 6444–6455.
Patel, T. P., et al. (2014). An open-source toolbox for automated phenotyping of mice in behavioral tasks. *Frontiers in Behavioral Neuroscience, 8*, 349.
Peixoto, L., Risso, D., Poplawski, S. G., Wimmer, M. E., Speed, T. P., Wood, M. A., et al. (2015). How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Research, 43*, 7664–7674.
Perry, C. J., & Lawrence, A. J. (2017). Hurdles in basic science translation. *Frontiers in Pharmacology, 8*, 478.
Pietropaolo, S., Singer, P., Feldon, J., & Yee, B. K. (2008). The postweaning social isolation in C57BL/6 mice: preferential vulnerability in the male sex. *Psychopharmacology (Berl), 197*, 613–628.
Pinter, O., Beda, Z., Csaba, Z., & Gerendai, I. (2007). Differences in the onset of puberty in selected inbred mouse strains. In *9th European Congress of Endocrinology* (pp. P617). Budapest, Hungary: Bioscientifica.
Puzzo, D., Lee, L., Palmeri, A., Calabrese, G., & Arancio, O. (2014). Behavioral assays with mouse models of Alzheimer's disease: Practical considerations and guidelines. *Biochemical Pharmacology, 88*, 450–467.
Pyter, L. M., Yang, L., McKenzie, C., da Rocha, J. M., Carter, C. S., Cheng, B., et al. (2014). Contrasting mechanisms by which social isolation and restraint impair healing in male mice. *Stress, 17*, 256–265.
Robbins, T. W. (2017). Cross-species studies of cognition relevant to drug discovery: a translational approach. *British Journal of Pharmacology*.
Robinson, L., & Riedel, G. (2014). Comparison of automated home-cage monitoring systems: Emphasis on feeding behaviour, activity and spatial learning following pharmacological interventions. *Journal of Neuroscience Methods, 234*, 13–25.
Ruby, N. F., Fernandez, F., Garrett, A., Klima, J., Zhang, P., Sapolsky, R., et al. (2013). Spatial memory and long-term object recognition are impaired by circadian arrhythmia and restored by the GABAAAntagonist pentylenetetrazole. *PLoS ONE, 8*, e72433.
Rutten, K., Reneerkens, O. A., Hamers, H., Sik, A., McGregor, I. S., Prickaerts, J., et al. (2008). Automated scoring of novel object recognition in rats. *Journal of Neuroscience Methods, 171*, 72–77.
Sakakibara, H., et al. (2012). Social isolation stress induces hepatic hypertrophy in C57BL/6J mice. *Journal of Toxicological Sciences, 37*, 1071–1076.
Sanchez, C., Asin, K. E., & Artigas, F. (2015). Vortioxetine, a novel antidepressant with multimodal activity: Review of preclinical and clinical data. *Pharmacology & Therapeutics, 145*, 43–57.
Sarter, M. (2006). Preclinical research into cognition enhancers. *Trends in Pharmacological Sciences, 27*, 602–608.
Schulz, J. B., Cookson, M. R., & Hausmann, L. (2016). The impact of fraudulent and irreproducible data to the translational research crisis – Solutions and implementation. *Journal of Neurochemistry, 139*(Suppl 2), 253–270.
Sik, A., van Nieuwehuyzen, P., Prickaerts, J., & Blokland, A. (2003). Performance of different mouse strains in an object recognition task. *Behavioural Brain Research, 147*, 49–54.
Silverman, J. L., Smith, D. G., Rizzo, S. J., Karras, M. N., Turner, S. M., Tolu, S. S., et al. (2012). Negative allosteric modulation of the mGluR5 receptor reduces repetitive behaviors and rescues social deficits in mouse models of autism. *Science Translational Medicine, 4* 131ra151.
Silverman, J. L., Oliver, C. F., Karras, M. N., Gastrell, P. T., & Crawley, J. N. (2013). AMPAKINE enhancement of social interaction in the BTBR mouse model of autism. *Neuropharmacology, 64*, 268–282.
Silvers, J. M., Harrod, S. B., Mactutus, C. F., & Booze, R. M. (2007). Automation of the novel object recognition task for use in adolescent rats. *Journal of Neuroscience Methods, 166*, 99–103.
Siuda, D., Wu, Z., Chen, Y., Guo, L., Linke, M., Zechner, U., et al. (2014). Social isolation-induced epigenetic changes in midbrain of adult mice. *Journal of Physiology and Pharmacology, 65*, 247–255.
Snyder, H. M., et al. (2016). Guidelines to improve animal study design and reproducibility for Alzheimer's disease and related dementias: For funders and researchers. *Alzheimers Dement, 12*, 1177–1185.
Sorge, R. E., et al. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods, 11*, 629–632.
Spooren, W., Lindemann, L., Ghosh, A., & Santarelli, L. (2012). Synapse dysfunction in autism: A molecular medicine approach to drug discovery in neurodevelopmental disorders. *Trends in Pharmacological Sciences, 33*, 669–684.
Stoppel, L. J., Kazdoba, T. M., Schaffler, M. D., Preza, A. R., Heynen, A., Crawley, J. N., et al. (2017). R-baclofen reverses cognitive deficits and improves social interactions in two lines of 16p11.2 deletion mice. *Neuropsychopharmacology*. http://dx.doi.org/10.1038/npp.2017.236 (in press).
Takahashi, Y., Sawa, K., & Okada, T. (2013). The diurnal variation of performance of the novel location recognition task in male rats. *Behavioural Brain Research, 256*, 488–493.
Talani, G., Biggio, F., Licheri, V., Locci, V., Biggio, G., & Sanna, E. (2016). Isolation rearing reduces neuronal excitability in dentate gyrus granule cells of adolescent C57BL/6J Mice: Role of GABAergic tonic currents and neurosteroids. *Frontiers in Cellular Neuroscience, 10*, 158.
van Goethem, N. P., Rutten, K., van der Staay, F. J., Jans, L. A., Akkerman, S., Steinbusch, H. W., et al. (2012). Object recognition testing: Rodent species, strains, housing conditions, and estrous cycle. *Behavioural Brain Research, 232*, 323–334.
Varty, G. B., Powell, S. B., Lehmann-Masten, V., Buell, M. R., & Geyer, M. A. (2006). Isolation rearing of mice induces deficits in prepulse inhibition of the startle response. *Behavioural Brain Research, 169*, 162–167.
Vijayanathan, V., Gulinello, M., Ali, N., & Cole, P. D. (2011). Persistent cognitive deficits, induced by intrathecal methotrexate, are associated with elevated CSF concentrations of excitotoxic glutamate analogs and can be reversed by an NMDA antagonist. *Behavioural Brain Research, 225*, 491–497.
Voelkl, B., & Wurbel, H. (2016). Reproducibility Crisis: Are we ignoring reaction norms? *Trends in Pharmacological Sciences, 37*, 509–510.
Vogel-Ciernia, A., & Wood, M. A. (2014). Examining object location and object recognition memory in mice. *Current Protocols in Neuroscience, 69*, 8–31.
Wahlsten, D., Metten, P., Phillips, T. J., Boehm, S. L., Burkhart-Kasch, S., Dorow, J., et al. (2003). Different data from different labs: Lessons from studies of gene-environment interaction. *Journal of Neurobiology, 54*, 283–311.
Wang, L., Jiao, J., & Dulawa, S. C. (2011). Infant maternal separation impairs adult cognitive performance in BALB/cJ mice. *Psychopharmacology (Berl), 216*, 207–218.
Wang, W., Pan, Y. W., Zou, J., Li, T., Abel, G. M., Palmiter, R. D., et al. (2014). Genetic activation of ERK5 MAP kinase enhances adult neurogenesis and extends hippocampus-dependent long-term memory. *Journal of Neuroscience, 34*, 2130–2147.
Wang, Y. Y., Smith, P., Murphy, M., & Cook, M. (2010). Global expression profiling in epileptogenesis: Does it add to the confusion? *Brain Pathology, 20*, 1–16.
Weiss, A., & Neuringer, A. (2012). Reinforced variability enhances object exploration in shy and bold rats. *Physiology & Behavior, 107*, 451–457.
Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., et al. (2015). Mapping sub-second structure in mouse behavior. *Neuron, 88*, 1121–1135.
Wohr, M., & Scattoni, M. L. (2013). Neurobiology of autism. *Behavioural Brain Research, 251*, 1–4.
Xiao, H., Liu, B., Chen, Y., & Zhang, J. (2016). Learning, memory and synaptic plasticity in hippocampus in rats exposed to sevoflurane. *International Journal of Developmental Neuroscience, 48*, 38–49.
Yang, M., Lewis, F. C., Sarvi, M. S., Foley, G. M., & Crawley, J. N. (2015). 16p11.2 Deletion mice display cognitive deficits in touchscreen learning and novelty recognition tasks. *Learning & Memory, 22*, 622–632.
Young-Pearse, T. L., & Morrow, E. M. (2016). Modeling developmental neuropsychiatric disorders with iPSC technology: Challenges and opportunities. *Current Opinion in Neurobiology, 36*, 66–73.
Yuan, R., Meng, Q., Nautiyal, J., Flurkey, K., Tsaih, S. W., Krier, R., et al. (2012). Genetic coregulation of age of female sexual maturation and lifespan through circulating IGF1 among inbred mouse strains. *Proceedings of the National Academy of Sciences of United States, 109*, 8224–8229.
Zhang, R., Xue, G., Wang, S., Zhang, L., Shi, C., & Xie, X. (2012). Novel object recognition as a facile behavior test for evaluating drug effects in AbetaPP/PS1 Alzheimer's disease mouse model. *Journal of Alzheimer's Disease, 31*, 801–812.
Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature, 465*, 690.