Title

Bad Behavior: Improving Reproducibility in Behavior Testing.

Journal

ACS Chemical Neuroscience, 9(8)

Authors

Andrews, Anne
Altieri, Stefanie
Yang, Hongyan
et al.

Publication Date

2018-08-15

DOI

10.1021/acschemneuro.7b00504

Peer reviewed

HHS Public Access

Author manuscript

ACS Chem Neurosci. Author manuscript; available in PMC 2018 August 20.

Published in final edited form as:

ACS Chem Neurosci. 2018 August 15; 9(8): 1904–1906. doi:10.1021/acschemneuro.7b00504.

Bad Behavior: Improving Reproducibility in Behavior Testing

Anne M. Andrews, **Xinyi Cheng**, **Stefanie C. Altieri**, and **Hongyan Yang**

Departments of Psychiatry & Biobehavioral Sciences and Chemistry & Biochemistry, Semel Institute for Neuroscience and Human Behavior, and Hatos Center for Neuropharmacology, University of California, Los Angeles

Abstract

Systems neuroscience research is increasingly possible through the use of integrated molecular, genetic, and circuit analyses. These studies depend on the use of animal models, and in many cases, behavioral analyses to investigate changes associated with genetic, pharmacologic, epigenetic, and other types of environmental manipulations. We illustrate typical pitfalls resulting from poor validation of behavior tests. We describe experimental designs and enumerate controls needed to improve reproducibility in investigating and reporting of behavioral phenotypes.

Reproducibility of study results is receiving increased scrutiny in biomedical research (*cf.*[1] and references therein). Inconsistencies in reported findings and irreproducible results are suggested to stem from failures in the design, execution, and analysis of experimental data.[1] Other factors contributing to poor reproducibility include inadequately characterized or poor quality reagents (*e.g.*, antibodies, cell lines, chemicals, animal strains), inadequate training and mentoring of personnel, complexities associated with collaborative studies, and a range of inappropriate responses to modern scientific pressures and incentives.[1]

In this Viewpoint, we focus on a common cause of failed reproducibility in research using animal models focused on behavioral outcomes—inadequate validation of laboratory specific testing conditions. Behavior experiments are designed to detect phenotypic differences between animals with varying genotypes, pharmacologic treatments, or environmental manipulations. Here, we outline the need for and steps associated with validating behavioral tests in individual laboratories upon first use and across time, including well-established paradigms. In doing so, we aim to provide authors, reviewers, and readers with guidelines for assessing the quality of behavior data and associated interpretations. We hypothesize that appropriately validated and controlled experiments will improve the reproducibility of behavioral findings across studies, research groups, and time.

As a first example, we recently added the novelty-suppressed feeding (NSF) test to a behavioral test battery used by our group (and others) to assess differences in anxiety-related behavior. The NSF test is used to detect hyponeophagia, *i.e.*, the ability of a novel environment to inhibit feeding behavior. The test relies on a conflict-avoidance paradigm where behavioral outcomes are balanced by competing demands between the natural tendency for rodents to avoid novel environments and the need to find food. Mice (or rats) are food deprived, which increases the incentive to feed. Each animal is then placed in the perimeter of a brightly lit, novel arena containing a food pellet. The latency to the first bite of food is recorded. Animals are also assessed in their home cages for latency to eat. Mice or

rats showing longer latencies to initiate feeding in the novel arena are regarded as exhibiting greater anxiety-related behavior than counterparts having shorter novelty-suppressed latencies to feed.

Using a strain of mice genetically engineered to lack serotonin transporter (SERT) expression[2] and previously reported NSF test parameters,[3] we were initially unable to observe increased anxiety-related behavior commonly associated with constitutive loss of SERT, as determined in the NSF and other anxiety-related behavior tests (Figure 1A).[2, 4–5] As such, we embarked on a systemic investigation aimed at varying key NSF test parameters. As shown in Figure 1B, we found that a longer food deprivation period (24 h *vs.* 18 h), warm white light (2700K *vs.* 5000K), intermittent light intensity (950 lux *vs.* 470 or 1200 lux), a smaller novel arena (19” W × 10” D × 8“ H *vs.* 20” W × 16“ D × 8” H) having a 2:1 ratio of dark to light walls and a larger, bright white food stage (6” W × 9” L *vs.* 4” W × 4” L), and testing 2 h after the light to dark switch were associated with a longer latency to feed in the novel arena in female and male mice with null SERT expression compared to wildtype siblings.

This example of NSF test validation illustrates two key points. First, had we proceeded with our experimental studies using the initial set of NSF test conditions (Figure 1A), we would have arrived at invalid or confounded conclusions when testing novel experimental groups. In other words, had we assumed the NSF test was “working” without first validating in our hands, the results of experiments investigating novel phenotypes could have been wrongly interpreted. We posit that by first reproducing generally accepted behavioral findings, researchers can increase confidence in the reproducibility of new findings. Second, we note that for behavior testing, and similar to other types of experiments, it is less critical to reproduce exact conditions reported by others, though these are a logical starting point. In contrast, it is more important to determine experimental conditions in individual laboratories that produce expected results, acknowledging that precise conditions vary across laboratories. In fact, trying to control for all variables (known and unknown) has not met with success in replicating behavioral findings.[6] *In summary, it is not the replication of precise experimental conditions that are of utmost importance but the ability to reproduce robust behavioral phenotypes under laboratory-specific conditions that are expected to improve reproducibility of novel behavioral phenotypes.*

We provide a second example to illustrate the necessity of behavioral test validation in Figure 2. In 2009, our research group moved from the Pennsylvania State University (PSU) to the University of California, Los Angeles (UCLA). The SERT deficient line of mice in Figure 1 was transferred *via* cryopreserved embryos. Following re-derivation at UCLA, we carried out experiments using the elevated plus maze (EPM), another common test for determining differences in anxiety-related behavior.[2] At first, we were unable to reproduce the increased anxiety-like EPM phenotype in mice lacking SERT, which we and others had reported repeatedly.[2, 4] This EPM phenotype is typically characterized by reduced open arm time and entries compared to wildtype mice (Figure 2).

In the case of the EPM, we were using the same strain of mice, the same maze, similar lighting conditions, *and* the same experimenter. While we were unable to identify the precise

reasons for the shift in the behavioral phenotype, we noted an overall trend toward greater open arm exploration at UCLA (from 20% to almost 50%, Figure 2). We concluded that the test conditions were not sufficiently averse at UCLA such that activity in the closed arms was equally likely *vs.* exploration of the open arms. Readjustment and validation of EPM test conditions, combined with using breeding parent-pairs selected to exhibit median anxiety-related characteristics, minimized phenotypic drift and "restored" the expected phenotype.

This second example highlights further observations. First, even within the same research group, behavioral phenotypes can shift and differences may go unnoticed without ongoing test validation. Behavior changes can be due to differences in laboratory or animal care personnel, environmental/housing conditions, or genetic or epigenetic drift. Thus, even laboratory-specific conditions benefit from periodic re-validation. Additionally, defining "expected" phenotypes raises a number of thorny questions. For example, in how many other laboratories, publications, or cohorts of animals from a single laboratory should a consistent phenotype be observed before a phenotype is deemed "expected"? What if after a reasonable period of validation involving systematic variations in test conditions, a laboratory is unable to reproduce a reported phenotype? When is it appropriate/important to report findings inconsistent with the literature? How robust does a phenotype need to be before it is worth studying? While the latter are difficult questions, their answers, in part, depend on instituting frequent validation of behavioral testing conditions.

In sum, we strongly advocate that data associated with new behavioral phenotypes need to be reported and interpreted in the context of well validated phenotypes. The latter can be evaluated *via* pretesting using animal models with well established behavioral outcomes or through the use of experimental groups integrated directly into study designs, *e.g.*, positive controls (Box 1). Baseline, positive, and internal controls for behavior studies bear resemblance to controls for "chemical" experiments (*e.g.*, internal standards, standard curves, signal-to-noise). Studies that include behavioral validation will add to the literature supporting reproducibility of existing behavioral phenotypes and improve the likelihood of reproducing newly reported phenotypes. Further, a single test by itself may be insufficient to capture complex behavioral phenotypes. As such, analyzing related variables across different tests, all of which need to validated, can help to elucidate novel phenotypes and to make valid comparisons with existing phenotypes.[2]

Authors are strongly encouraged to include behavioral test validation in their experimental designs. Reviewers and editors are advised to request behavior test validation as part of the review process. And readers should expect validation so as to interpret novel findings with confidence and to anticipate their reproducibility over time and across studies.

Acknowledgements:

The authors gratefully acknowledge Hye Rhyn (Hannah) Chung, Brandon Yoshida, and Westin E. Babyak for assistance with behavior experiments, Dr. Jacqueline Crawley for feedback on the manuscript, and the National Institute of Mental Health for financial support (MH086108 and MH106806).

Literature Cited

1. Flier JS, Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy. Mol Metab 2017, 6 (1), 2–9. [PubMed: 28123930]
2. Altieri SC; Yang H; O'Brien HJ; Redwine HM; Senturk D; Hensler JG; Andrews AM, Perinatal *vs.* genetic programming of adult serotonin states associated with anxiety. Neuropsychopharmacology 2015, 40 (6), 1456–1470. [PubMed: 25523893]
3. Samuels BA; Hen R, Novelty-suppressed feeding in the mouse In Mood and Anxiety Related Phenotypes in Mice: Characterization Using Behavioral Tests, Gould TD, Ed. Springer: 2011; Vol. II.
4. Holmes A; Li Q; Murphy DL; Gold E; Crawley JN, Abnormal anxiety-related behavior in serotonin transporter null mutant mice: The influence of genetic background. Genes Brain Behav 2003, 2 (6), 365–80. [PubMed: 14653308]
5. Ansorge MS; Zhou M; Lira A; Hen R; Gingrich JA, Early-life blockade of the 5-HT transporter alters emotional behavior in adult mice. Science 2004, 306 (5697), 879–81. [PubMed: 15514160]
6. Crabbe JC; Wahlsten D; Dudek BC, Genetics of mouse behavior: interactions with laboratory environment. Science 1999, 284 (5420), 1670–2. [PubMed: 10356397]

Box 1:

Controls for Behavior Test Validation

- ***Baseline control groups*** (wildtype or vehicle-treated mice) are used to determine whether a test "works" under laboratory-specific conditions (*e.g.*, novel arena suppression of latency to feed *vs.* home cage environment (NSF), reduced open arm compared to closed arm exploration (EPM)).
- ***Positive control groups*** are used to determine whether a test detects "expected"/previously-reported phenotypic changes (*e.g.*, SERT-deficient mice show greater latency to feed in the novel arena (NSF) and reduced open arm activity (EPM) *vs.* wildtype mice.) Positive control groups are also used to assess dynamic range, (*i.e.*, to detect "floor" or "ceiling" effects that could interfere with observing changes in a novel test group).
- ***Internal controls*** are integrated into test designs and help to determine test validity (*e.g.*, comparisons to determine if baseline and positive control groups exhibit similar latencies to feed in their home cages (NSF), or closed arm time/entries as measures of locomotor activity (EPM)).

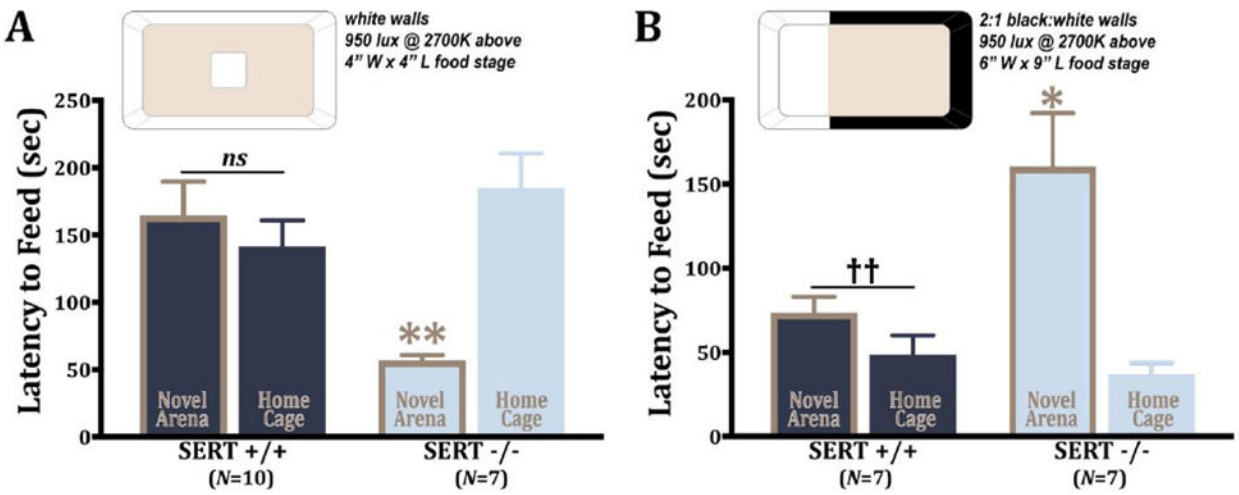

Figure 1. Validation of novelty-suppressed feeding test conditions.

Three-five-month-old serotonin transporter (SERT)-deficient mice were evaluated under different experimental conditions and types of novel arenas. Wildtype (SERT+/+) mice served as the baseline control group, while SERT−/− subjects served as the positive control group for test-condition validation. Three criteria were used to determine the validity of testing conditions: similar latencies to feed between baseline and positive control groups in the home cage, increased latency to feed in the novel arena relative to the home cage in SERT+/+ mice, and potentiation of increased latency to feed in the novel arena in SERT−/− *vs.* SERT+/+ groups. Only experimental conditions in (B) fulfilled all criteria and were considered valid. Testing conditions in (A) were not valid because (1) the baseline group (SERT+/+) failed to distinguish the novel arena from the home cage with an increased latency to feed, though the latency to feed in the home cage was similar across SERT+/+ and SERT−/− groups. And (2), in (A), the positive control group (SERT−/−) showed a shorter and not longer latency to feed in the novel arena *vs.* the home cage and compared to SERT +/+ mice in the novel arena. Data are means ± SEMs. **P*<0.05 and ***P*<0.01 are *t*-test comparisonswith respect to SERT+/+ mice in the novel arena; ††*P*<0.01 is a paired *t*-test between SERT+/+ mice in the novel arena *vs.* home cage.

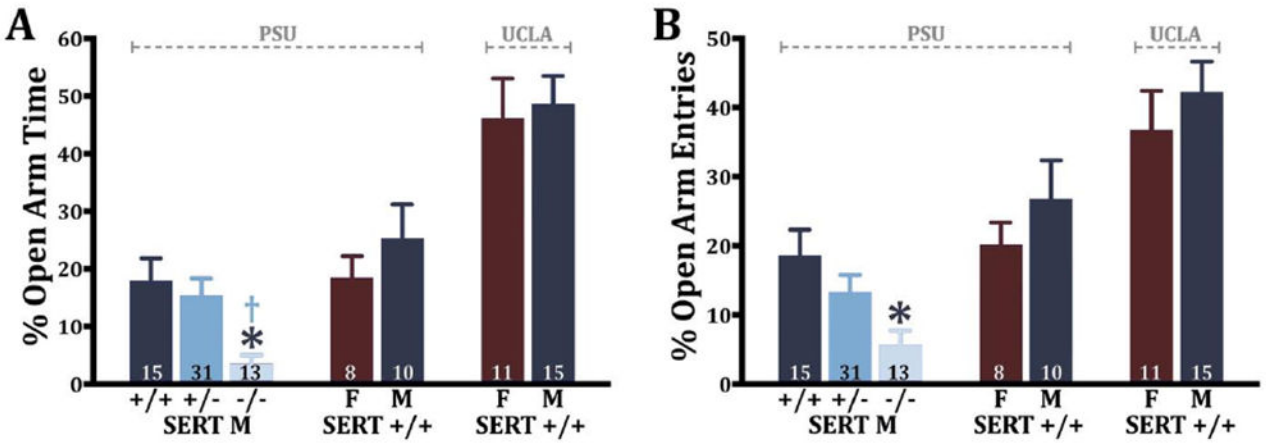

Figure 2. Validation of elevated plus maze test conditions.

Individual cohorts of two-three-month-old serotonin transporter (SERT)-deficient female (F) and male (M) mice were tested at the Pennsylvania State University (PSU) and the University of California, Los Angeles (UCLA). In each case, wildtype (SERT+/+) mice served as the baseline controls and SERT−/− subjects constitued the positive control groups. Two internal controls were used to examine the validity of the testing conditions: (A) %open arm time with respect to total time in the open and closed arms, and (B) %open arm entries with respect to total entries in the open and closed arms. Test conditions were initially validated at PSU, based on an expectation for reduced activity in the open arms relative to the closed arms in the baseline group (SERT+/+) (*i.e.*, <50% of total arm time in the open arms (A, middle set of bars) and <50% of total arm entries in the open arms (B, middle bars)). Moreover, the positive control group (SERT−/− mice) exhibited an increased anxiety-related phenotype evidenced by lower open arm activity relative to the baseline control group (SERT+/+ mice). However, the test conditions failed to validate shortly after relocating to UCLA because neither SERT+/+ (baseline control) nor SERT−/− mice (positive control) groups distinguished the anxiogenic open arms from the closed arms (*i.e.*, %open arm time and %open arm entries were close to 50%). Group sizes are indicated at the bottom of each bar. Data are means ± SEMs. **P*<0.05 *vs.* SERT+/+ male mice; †P<0.05 *vs.* SERT+/− (SERT heterozygous male mice.