Measuring behavior of animal models: faults and remedies

Ehud Fonio, Ilan Golani & Yoav Benjamini

Widely used behavioral assays need re-evaluation and validation against their intended use. We focus here on measures of chronic anxiety in mouse models and posit that widely used assays such as the open-field test are performed at the wrong time, for inadequate durations and using inappropriate mouse strains. We propose that behavioral assays be screened for usefulness on the basis of their replicability across laboratories.

An editorial published in *Nature Methods*[1] pointed out that as more mouse models are produced, researchers studying neuropsychiatric diseases will need better ways to evaluate them. The editorial asserted that it is essential to discuss the strengths and weaknesses of animal models, develop new attitudes toward the measurement of behavior and design new and more complex tests of behavior that are up to the task they claim to fulfill. The editorial joins a long list of reviews criticizing available animal models of behavior that call for better experimental setups, mouse models and measures[2–6].

The validity of the open-field test (OFT), for example, used since 1934 as a tool for assessing emotionality[7], has been questioned since 1973 (ref. 8). In 1976, an extensive review analyzed the variability in the ways different laboratories were performing the test[9]. The study showed that the physical features of the OFT arena—such as size, materials, colors and shapes—and the parameters being measured were extremely diverse across laboratories. Moreover, OFT studies rarely included reports of procedural detail with regard to the conditions that the animals had experienced prior to the experiments, and the studies varied greatly with regard to the analyses and interpretations being made. OFT field studies were characterized by limited reliability, poor validation of measures and difficulty in interpreting these measures. Given these weaknesses, the authors of that review concluded that "It is small wonder that contradictions and failures of replication abound in OFT research"[9].

Open-field studies have indeed been yielding inconsistent and even opposing results up to this time[10]. But lack of standardization is not the sole cause of the problem: a mouse phenotyping study performed simultaneously in three laboratories made excessive efforts to standardize the setup and experimental conditions across laboratories, but its authors nonetheless concluded that the results of OFTs were often idiosyncratic to particular laboratories[11]. This state of the art also characterizes other common tests of behavior, in particular those used for the estimation of anxiety. One review lists some 30 anxiety tests that require urgent attention to the much-neglected issue of behavioral validation. The most popular anxiety test in this list is the elevated plus maze (EPM). The author of that review, himself an extensive user of this test, attributes its popularity to practical rather than theoretical considerations[3]. Reviews focused on other tests do not portray a more positive profile[4,6,12,13].

Aside from its impact on basic neuroscience studies, this problem also affects the field of drug discovery, which has a worldwide economic cost of over $40 billion per year[6,14]. In this field, the relative failure of current animal models (including experimental setups, mouse models and measures) to predict the effect of drugs on human disease of the central nervous system is attributed to a multitude of hypothetical reasons ranging from molecular to behavioral to sociological[15].

Regardless of the above, the number of publications using the OFT has increased fourfold in the last 10 years[10]. Furthermore, the weaknesses reported in the reviews mentioned above have hardly been amended[13], and efforts to improve the quality of measured behavior[16] have not been widely embraced. What could be the explanation for this continued failure of the common tests, which, instead of showing diminished use, show a gain in popularity?

As pointed out in the editorial[1], "the accumulation of data is not enough. Appropriate attitudes are also required. Researchers too often simply assert or assume the validity of their models and assays…. The validity of a particular model depends on the goal…."

Although a review of the rich tradition of discussing the types of validity of models, assays and measures is beyond the scope of this commentary, here we examine those aspects of validity of behavioral measures that could be readily addressed by researchers. When using a behavioral test, researchers should first ask: are we measuring what we intend to measure with it? Obviously, a requirement of any assay is that it will reflect essential features of the behavior it purports to model. Once the answer is positive, the next question is: are we measuring this behavior in a useful way? In other words, is the measure reliable?—that is, is it stable

Ehud Fonio is in the Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel. Ilan Golani is in the Department of Zoology, The George S. Wise Faculty of Life Sciences the Sagol School of Neuroscience, and Yoav Benjamini is in the Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences and the Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel.
e-mail: ehud.fonio@weizmann.ac.il

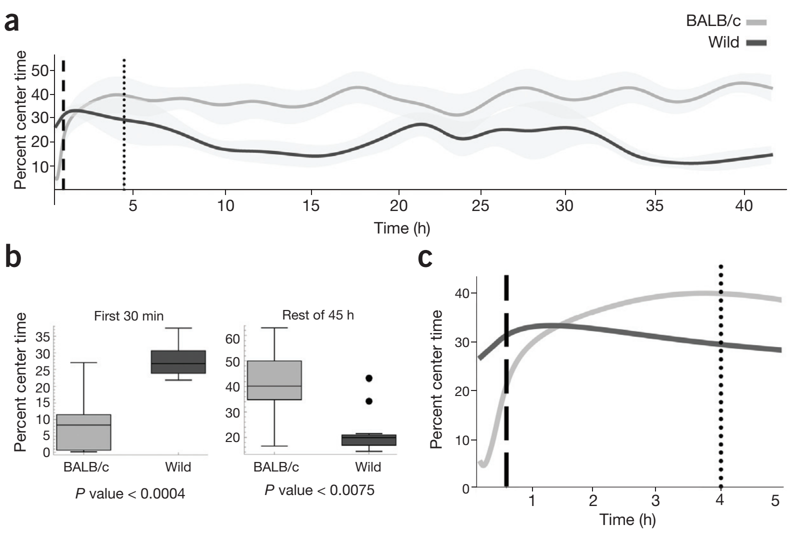

Figure 1 | Comparing anxiety behavior between BALB/c and wild mice. (**a**) 'Percent center time', a classical measure of anxiety, plotted over a 45-h period. Curve represents mean ± s.e.m. (gray shading). Dashed vertical line demarcates the end of the first 0.5 h; dotted line demarcates the end of the transient habituation period, which is estimated quantitatively[15]. (**b**) Quantification of percent center time values for each mouse strain in the first 30 min and the rest of the session. Box plot summaries compare the respective values in the first half hour and the rest of the session. The bottom and top of the box are the lower and upper quartiles, respectively, and the horizontal bar near the middle of the box is the median; the ends of the whiskers represent the lowest and highest data points within an interquartile range of 1.5 from the first and third quartiles, respectively. Outliers are represented as dots. (**c**) Percent center time along the first 5 h in BALB/c and wild mice; vertical lines as in **a**. BALB/c, *n* = 12; wild mice, *n* = 9.

for a single animal, for a group of animals and across groups of animals from different laboratories? If we can affirmatively answer both questions, most of the above objections to the use of common simple measures and tests should disappear.

Alas, for many widely used behavioral tests, the answer to even the first question is negative. Let us demonstrate this point with an example obtained from our own recent study, which used a modified OFT to measure behavior in a mouse model of chronic anxiety[15]. An animal model of generalized anxiety disorder should, by definition, demonstrate long-lasting and stable characteristics of anxious behavior that should persist over long time periods without habituation to the environment[2,17]. To examine whether these conditions are fulfilled for the OFT, in a recent study[15] we analyzed 'percent center time' occupancy in an open-field setup that included a home cage allowing free passage to a large (2.5-m-diameter) circular arena[18,19] and analyzed animal behavior over 45 h, which is notably longer than the maximal half-hour period used in the vast majority of common studies of anxiety (**Fig. 1**). In this setup, we compared the behavior of the prototype mouse strain used as an animal model of anxiety, BALB/c mice, to first-generation-in-captivity wild mice, which had already been used in a previous study as an ethologically relevant reference[20]. BALB/c mice exhibited a transient period of high anxiety, marked by low percent center time, followed by a long-lasting stable period of calm behavior. In contrast, wild mice exhibited a transient low-anxiety period followed by consistent anxious behavior relative to the other strain. BALB/c mice thus scored significantly higher on anxiety (lower percent of center time occupancy) over the first 30 min, whereas wild mice scored significantly higher on anxiety over the rest of the 45-h session. Thus, at least with this measure, the first 30 min did not reflect the general behavior of the animals; rather, a transient period of habituation preceded a long stable period characterized by results opposite to those obtained in the transient period. Notably, the observed reversal in the behavior of the two strains took place after the half-hour period, which is the maximal period analyzed in the vast majority of anxiety studies. (It is not uncommon for EPM, OFT and light-dark box tests of behavior to be measured for even shorter durations, commonly 5 min or 10 min). Many studies use the test-retest procedure, testing the animal more than once for the same short duration, usually a few days apart, to assess temporal stability of the behavior[17]; but in our view this is not a remedy for the short test duration and merely results in taking the inappropriate measurements twice.

All other measures used in our study, which are equivalent to corresponding measures in the OFT, EPM and light-dark box tests, behave in much the same way as percent center time[15]. These include 'percent of time spent in the open area', 'percent of arrest (freezing) time', 'arrest duration', 'number of transitions' between the home cage and the arena, and 'activity' (distance traveled in arena). Notably, for all measures, a computable transient period of habituation preceded the long stable period that captured results opposite to those obtained in the transient period[15].

Models of 'pathological' anxiety (such as generalized anxiety disorder) are often referred to as 'trait' anxiety tests. Trait anxiety is, by definition, the persistent and durable feature of an individual's personality that reflects the way one interacts with one's physical and social environment[2,17]. Unlike 'state' anxiety, trait anxiety does not vary from moment to moment and is considered to be an enduring feature of an individual[12]. As our study shows, behavioral measurement of BALB/c mice in the OFT over short time periods reflects a temporary state that is not reflective of chronic anxiety behavior and thus fails to fulfill the requirement for a model of chronic, long-lasting anxiety[2–6]. Furthermore, use of cumulative statistical measures during the first half hour, which involves a consistent and large change probably reflecting habituation to the novel area, misses the dynamics of growth characterizing this transient[18,19].

By extending test duration, we show that one can study both transient and enduring properties of the behavior in the same setup. By comparing the behaviors of domesticated to wild *Mus musculus*, we show that following a habituation period, the default of the domesticated strain is calm behavior, whereas the default of wild mice is anxious behavior; therefore, wild mice appear to be a better model for chronic anxious behavior.

We propose that chronic anxiety should be captured in a prototype similar to that exhibited by wild mice from the end of the fourth hour (after the introduction of the mouse to the open field) and during a 4-h interval. Indeed, the validity of this modified assay still relies on the assumption that the measures listed above reflect anxiety. To fully establish this, it would be necessary to examine the behaviors of wild-derived animal models under intact or manipulated situations, as has been recently suggested[21]. This can be accomplished by using pharmacology (anxiolytic and anxiogenic drugs)[22], by using genetic modification, by studying the effects of environmental manipulations that change the animals' stress level, and by examining wild-type behavior in other domains not involving exploration.

We thus conclude that, at least in the context of chronic diseases, to answer the question of relevance we should first 'zoom out' to obtain a wider perspective that will help us search for better experimental setups, strains and measurement procedures.

There is no question that scientists are in need of a standard method that will allow them to present their results. The reasons why behavioral assays of anxiety have been done in the wrong way for such a long time reflect—in our view—a combination of the following: the convenience of using a test of short duration that allows automated high-throughput experimentation; the difficulty in publishing results obtained with new, nonstandard methods; and an insufficient interest from researchers in the problem of adequate measurement or, for that matter, in the structure of behavior (which is, in turn, essential for deciphering the meaning of behavior).

With respect to the second issue, 'Will measuring behavior over longer time periods be useful?': inspection of the measure variability across the group as a function of the duration over which center-time occupancy is measured can answer this question. According to our estimations, center occupancy and the other five measures listed above should be measured for a few hours for the group variance to be small enough.

But for a behavioral measure to be useful, it is not sufficient for it to merely have a small variance over a group of mice tested under the same condition in a single laboratory. Rather, effect differences (such as strain differences or mutants versus background differences) evaluated in different laboratories should also yield similar results. In other words, the measure has to demonstrate replicability across laboratories.

The ongoing practice in the field to achieve cross-laboratory replicability is to increase the level of standardization of the test's protocol and of the environmental factors involved, such as raising, feeding, housing and handling of animals[23,24]. An alternative approach that we first proposed in Kafkafi *et al.*[25] is to accept the fact that laboratories are different in unpredictable ways and that no level of standardization can entirely avoid this because there is no way to standardize environmental features of which one is unaware. In Kafkafi *et al.* we thus proposed that experiments that validate behavioral measures should be carried out in more than one laboratory, with no out-of-the-ordinary standardization, and statistical estimates of the variability of the interaction between genotype and laboratory for the measure should be obtained. Such so-called interaction variability occurs when, for example, the experimenter does not know that one of the strains is blind and that the lights are brighter in one of the laboratories, thus affecting in that laboratory only the sighted strain.

Such situations cause more variation than is typically observable in a single laboratory. We believe, therefore, that insisting on candidate measures that are stable across groups of animals in different laboratories as well as relevant for individual animals and stable for groups of animals in the same laboratory will aid the development of behavioral assays and measures that reflect meaningful differences.

Should all research making use of behavioral tests be conducted in multiple laboratories? Not necessarily. In Richter *et al.*[26], the authors demonstrated how deliberate introduction of adequate environmental heterogeneity into the design of a single-laboratory experiment generates variability that reflects the variability inherent in multilaboratory studies that do not ignore the interactions in their analysis.

An alternative option would be for developers and users of behavioral measures to cooperate and create large phenotyping databases. A collection of data related to behavioral measurements of individual animals acquired in multiple laboratories, at multiple times, involving multiple strains, and with no out-of-the-way standardization efforts could be a resource of great value for evaluating the validity of different behavioral measures. Such databases have been established in recent years: for example, the EuroPhenome database[27] (http://www.europhenome.org/), the WebQTL's Published Phenotypes database[28] (http://www.genenetwork.org/) and the Mouse Phenome Database[29] (http://phenome.jax.org/). Although these as yet include only a limited number of laboratories and genotypes, they all try to enlist larger groups of researchers and to expand the animal models covered, and they are publicly available. It will be beneficial for the redesign of new behavioral measures that raw behavioral data will be available as well in these databases.

Access to this information will allow experimenters to extract from the database the size of the genotype-by-laboratory interaction relevant to their experiment. The experimenters can then conduct their work in their own laboratory and combine their in-lab variability with the outside information on interaction variability, which will help them obtain more realistic estimates of variance[25]. It is reassuring to observe the coordinated efforts going into the construction of the database, but more effort is required to develop the analysis tools needed for the use of the databases for the above purpose.

Both this proposal and the approach of Richter *et al.*[26] rely on the conviction that effects demonstrated against the higher yardstick of variability, one that captures the size of genotype-by-laboratory interactions, are likely to be replicable in another laboratory.

In agreement with the editorial[1], it is our experience that even widely used measures of behavior have to be re-evaluated and validated against their intended use. Their usefulness has to be further assessed by performing the same experiment in multiple laboratories. Once the validity and usefulness are established, the measures can be used with the information mined from the current databases, even in single-laboratory experiments. Alas, if their validity or their usefulness for a particular goal cannot be established, there is no alternative but to return to the design table and investigate the behavior in detail to come up with new and better measures[16].

ACKNOWLEDGMENTS

Our research is supported by grants from the Israel Science Foundation of The Israel Academy of Science to I.G. and Y.B. and by a special grant from Tel Aviv University's vice president of research. Y.B. was supported at the later stage of the research by European Research Council Advanced Researcher grant (PSARPS). We thank Noldus EthoVision for letting us use their tracking system and N. Kafkafi for his assistance in writing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available at http://www.nature.com/doifinder/10.1038/nmeth.2252.

1. Anonymous. *Nat. Methods* **8**, 697 (2011).
2. Lister, R.G. *Pharmacol. Ther.* **46**, 321–340 (1990).
3. Rodgers, R.J. *Behav. Pharmacol.* **8**, 477–496 (1997).
4. Ennaceur, A., Michalikova, S., van Rensburg, R. & Chazot, P.L. *Behav. Brain Res.* **188**, 136–153 (2008).
5. Markou, A., Chiamulera, C., Geyer, M.A., Tricklebank, M. & Steckler, T. *Neuropsychopharmacology* **34**, 74–89 (2009).
6. Nestler, E.J. & Hyman, S.E. *Nat. Neurosci.* **13**, 1161–1169 (2010).
7. Hall, C.S. *J. Comp. Psychol.* **18**, 385–403 (1934).
8. Archer, J. *Anim. Behav.* **21**, 205–235 (1973).
9. Walsh, R.N. & Cummins, R.A. *Psychol. Bull.* **83**, 482–504 (1976).
10. Stanford, S.C. *J. Psychopharmacol.* **21**, 134–135 (2007).
11. Crabbe, J.C., Wahlsten, D. & Dudek, B.C. *Science* **284**, 1670–1672 (1999).
12. Kalueff, A.V., Wheaton, M. & Murphy, D.L. *Behav. Brain Res.* **179**, 1–18 (2007).
13. Bourin, M., Petit-Demoulière, B., Dhonnchadha, B.N. & Hascöet, M. *Fundam. Clin. Pharmacol.* **21**, 567–574 (2007).
14. Kola, I. & Landis, J. *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
15. Fonio, E., Benjamini, Y. & Golani, I. *PloS ONE* **7**, e48414 (2012).
16. Benjamini, Y. *et al. Neurosci. Biobehav. Rev.* **34**, 1351–1365 (2010).
17. Andreatini, R. & Bacellar, L.F.S. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **24**, 549–560 (2000).
18. Fonio, E., Benjamini, Y. & Golani, I. *Proc. Natl. Acad. Sci. USA* **106**, 21335-21340 (2009).
19. Benjamini, Y., Fonio, E., Galili, T., Havkin, G.Z. & Golani, I. *Proc. Natl. Acad. Sci. USA* **108**, 15580–15587 (2011).
20. Fonio, E., Benjamini, Y., Sakov, A. & Golani, I. *Genes Brain Behav*. **5**, 380–388 (2006).
21. Koide, T. *et al. Exp. Anim.* **60**, 347–354 (2011).
22. Jain, A., Dvorkin, A., Fonio, E., Golani, I. & Gross, C.T. *Eur. Neuropsychopharmacol.* **22**, 153–163 (2012).
23. Baker, M. *Nature* **475,** 123–128 (2011).
24. Williams, S.C.P. *Nat. Med.* **17**, 1324 (2011).
25. Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G.I. & Golani, I. *Proc. Natl. Acad. Sci. USA* **102**, 4619–4624 (2005).
26. Richter, S.H., Garner, J.P., Auer, C., Kunert, J. & Würbel, H. *Nat. Methods* **7**, 167–168 (2010).
27. Morgan, H. *et al. Nucleic Acids Res.*, **38** (suppl. 1), D577–D585 (2010).
28. Chesler, E.J., Lu, L., Wang, J., Williams, R.W. & Manly, K.F. *Nat. Neurosci.* **7**, 485–486 (2004).
29. Grubb, S.C., Churchill, G.A. & Bogue, M.A. *Bioinformatics* **20**, 2857–2859 (2004).