Journal of Neuroscience Methods 234 (2014) 2–12

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Basic Neuroscience

Reproducibility and relevance of future behavioral sciences should benefit from a cross fertilization of past recommendations and today's technology: "Back to the future"

Berry M. Spruijt[a,b,*], Suzanne M. Peters[a,b], Raymond C. de Heer[a,b], Helen H.J. Pothuizen[b], Johanneke E. van der Harst[a,b]

[a] Department of Biology, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
[b] Delta Phenomics B.V., Nistelrooisebaan 3, 3574 RE Schaijk, The Netherlands

HIGHLIGHTS

- Standard behavioral tests have limited validity and lack reproducibility.
- Advanced methods and technologies are available, but marginally applied.
- The lack of using innovative test approaches hampers behavioral sciences.
- Here we discuss this 'status quo' and propose solutions to advance behavioral testing.
- Examples are given to illustrate the enlarged discriminability of novel test methods.

ARTICLE INFO

Article history:
Received 21 November 2013
Received in revised form 3 March 2014
Accepted 4 March 2014
Available online 12 March 2014

Keywords:
Animal behavior
Open field
Validity
Automation
Home-cage testing
Ethology

ABSTRACT

Thanks to the discovery of novel technologies and sophisticated analysis tools we can now 'see' molecules, genes and even patterns of gene expression, which have resulted in major advances in many areas of biology. Recently, similar technologies have been developed for behavioral studies. However, the wide implementation of such technological progress in behavioral research remains behind, as if there are inhibiting factors for accepting and adopting available innovations. The methods of the majority of studies measuring and interpreting behavior of laboratory animals seem to have frozen in time somewhere in the last century. As an example of the so-called classical tests, we will present the history and shortcomings of one of the most frequently used tests, the *open field*. Similar objections and critical remarks, however, can be made with regard to the elevated plus maze, light-dark box, various other mazes, object recognition tests, etc. Possible solutions and recommendations on how progress in behavioral neuroscience can be achieved and accelerated will be discussed in the second part of this review.

Contents

* Corresponding author at: Department of Biology, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. Tel.: +31 30 2535364; fax: +31 30 2532837.
E-mail addresses: b.m.spruijt@uu.nl, nancy.paalman@deltaphenomics.com (B.M. Spruijt).

http://dx.doi.org/10.1016/j.jneumeth.2014.03.001
0165-0270/

1. General introduction

The first open field study was conducted in 1934 (Hall, 1934, 1936). Next to its use for measuring locomotor activity, the open field often has been used for studies on anxiety. Effects of strains, gender, age, early life experience, enrichment, illumination, all kind of odors, temporal aspects of other dependent parameters on emotionality were already mentioned almost 40 years ago (Walsh and Cummins, 1976), see Fig. 1. The recommendations forwarded by Walsh and Cummins (1976), however, seem to have been ignored. The use of this test has since then rapidly increased, but in a more simplified way than was proposed in 1976. Thus, it is not surprising that despite this early recommendation to improve the test, Haller and Alicki recently reported the low validity of this test (and other tests) in anxiety research (Haller and Alicki, 2012). One might expect that its use, the protocols applied and the interpretation of parameters are well validated delivering reliable and reproducible data. Yet, unfortunately, this was not the case in 1976 and is still not in 2013.

The lack of reliability and validity of so-called 'classical standard tests' (like the open field, elevated plus maze, etc.) have been repeatedly addressed (Benjamini et al., 2010; Crabbe et al., 1999; Kafkafi et al., 2005; McClearn, 2004; Mandillo et al., 2008). As indicated in the paper of Richter (Richter et al., 2011), the notion that there is lack of reproducibility of results between studies seems to be widespread (see for example across laboratory data compared by Wahlsten et al., 2003). Differences in sensitivity to drugs of abuse seen in repeated testing of various inbred strains have been reported (Cabib et al., 2000) as well as across laboratory variability in thermal nociception (Chesler et al., 2002). In a retrospective study comparing results of decades of research on behavior involving anxiety and agonistic behavior (Wahlsten et al., 2006), it was concluded that only alcohol preference and general locomotor activity (expressed as centimeters per minute) in an open field test lasting maximally 15 min, are reliable parameters (see also Wahlsten et al., 2003). In contrast to the aforementioned supposed robustness of the locomotion readout, it was demonstrated in 2004 that the most frequently used readouts: i.e. distance traveled and time spent in the center, explained only 9.6% of the path variability (Lipkind et al., 2004). Nonetheless, Wahlsten and co-workers consider their findings as sufficient evidence for the robustness of this test, ignoring its short duration and the limited ethogram and, thus, ignoring the recommendations forwarded in the review of 1976 and the critical notes raised in 2004.

The unexplained variability of behavioral variables can be assigned to all kind of influences. We demonstrate here that habituation is a process of hours and days, and not only of minutes. This is exemplified in Fig. 2 which shows the continuous monitoring of five inbred strains during the active (dark) phase of the day. Activity declines over the days and the relative differences between the strains changes: C3H mice are more active than DBA/2 mice from the fifth day onwards, which is in clear contrast to the first day when they are less active. Apparently, there are different types of habituation: (i) the relatively fast diminishment in activity across hours, typically seen in short lasting tests and (ii) more baseline levels of activity seen after days. The first probably involves a quick reaction to novelty to acquire knowledge on spatial cues and most relevant stimuli of the environment. The second phase of baseline activity develops when the predictability and safety of the environment have been assessed for instance by the regular access to food, water and a safe place. The almost 50% reduction in activity over days suggests that animals start to behave more routinely in a predictable environment.

Thus, virtually all open field studies (which typically last five to 60 min) only measure immediate reactions to novelty. Moreover, activity measured as distance traveled consists of a number of distinct components, such as length, speed, angularity, number of stops, etc. (see Fig. 2). A more precise distinction and measurement of these variables enhances the discriminability and the variability. Pfaff (2001), as cited in Wahlsten et al. (2006), labeled variability as an unstable feature of behavior. Here, we consider this variability as phenotypic plasticity, as a core feature of behavior, which should be acknowledged and pursued in analyses for an improved reproducibility and a more precise characterization of behavior. Also Benjamini and colleagues (2010) outlined the added value of a more detailed definition of behavior. They use frequency distributions of specific endpoints of each individual animal to define the transitions between behavioral elements of that animal, rather than using arbitrarily chosen criteria similar for all animals. Thus, more "animal centered measures", sufficient time and a larger environment or arena to explore may contribute to a larger reproducibility and discriminability of exploration.

Shortcomings of behavioral studies have been highlighted throughout the literature, emphasizing the need for more ethological procedures (Fonio et al., 2012; Gerlai, 2001, 2002; Gerlai and Clayton, 1999; Würbel, 2002, 2009) and improved methods of analysis (Benjamini et al., 2010; Fonio et al., 2009; Golani, 2012). The need for continuous animal observations in a more social and natural setting have led to the development of automated tools (Ohayon et al., 2013; Weissbrod et al., 2013). Ohayon et al. (2013) have developed an automated system that permits the reproducible continuous monitoring of social behavior of mice for five days. Thus, shortcomings of classical standard tests are known and innovations are made. Yet, they are neither widely accepted, nor adopted (Haller