Detecting, measuring and correcting errors in automated behavior analysis equipment

Richard E. Brown
Psychology Department, Dalhousie Universit, Halifax,
Nova Scotia, Canada B3H 4J1, rebrown@dal.ca

Behavioral phenotyping of transgenic and mutant mice depends on the observation and recording of behavior of specific strains of mice in a particular test apparatus in a particular laboratory [2]. Not all researchers, however, obtain the same results when the same strains of mice are tested in the same apparatus in different laboratories [11]. Chesler et al. [4] calculated that experimenter effects were responsible for a significant proportion of the laboratory effects in behavioral research. These experimenter effects include observer bias, observational errors, and recording errors [7]. In order to reduce experimenter error and conduct high throughput analysis of behavior, many different types of automated behavioral test apparatus have been developed [10].

Automated activity recorders, video-tracking systems, and scoring equipment has been designed for a number of behavioral tests, including the open field, Morris water maze, fear conditioning, and home cage activity. When a researcher uses such automated equipment, how do they know that the data collected are valid and reliable? How can the accuracy of automated equipment be measured?

One method is to compare the scores of automated equipment with those of trained observers. Fitch et al. [5] compared their automated movement detector for fear conditioning with the results of trained observers using three different scoring methods to record freezing behavior. Although all four methods gave comparable results, the automated recording apparatus seemed less sensitive to freezing than event-recorder or time-sampling methods, and event recording allowed the observer to quantify other behaviors. Although Fitch et al. [5] favour of the automated system, the event-recorder and time-sampling methods may provide more information and can be just as accurate. The problem is that they are more time consuming. Crowley et al. [3] found that the scores from automated forced swim and tail suspension tests correlated highly with the scores of three observers but argue that the cost-benefit analysis indicates an advantage of the automated system.

In their analysis of freezing behavior in fear conditioning, Marchand et al. [8] found high correlations (r= 0.995) between observers and between observers and the SUB automated scoring system. However, the two automated scoring methods (SUB vs RAW) produced different results and made different types of errors, classifying some bouts of walking, rearing, sniffing, grooming and moving behavior as "freezing". Automated equipment is also used to detect more complex behaviors. Graziano et al. [6] calculated that their automated system correctly categorized 97.9% of swim path types in the Morris water maze, as determined by three human observers, with 22 errors in 1049 swim path analyses. Nadler et al. [9] found that two trained observers had a 95% agreement and did not differ in their scores from the automated equipment in scoring sociability (time spent near a target mouse) and social novelty (time spent near a familiar vs. a novel mouse).

In our research, we have compared the results from trained observers and videotape analyses with those of automated equipment and found a number of discrepancies between the data from the automated equipment and the observers. I shall discuss six examples:

Automated open fields give higher activity scores than observer-based open fields. In testing Coloboma mice we found that the automated open field gave very high "horizontal activity" scores because circling was scored as horizontal activity and not as stereotyped behavior.

Automated open fields track only part of the mouse, not the whole mouse. When scoring activity, we score a movement only when all four feet of the mouse cross a line, but our automated system scored only the front half of the mouse, thus transition scores were inflated.

The automated Barnes maze made errors in scoring"head pokes" into holes. Our tracking system defined a zone around the hole as an "error zone" and scored an error when a mouse entered this zone, even if it did not head-poke into the hole. Thus, the automated system scored far more errors than the human observers.

Automated tracking system errors. We found tracking system errors on the elevated plus maze, open-field and Morris water maze as the tracking system recorded behavior "outside the apparatus" and showed the mouse travelling through the walls of the apparatus or "flying" from one arm of the elevated plus maze to another. We corrected these errors manually. They appear to be due to miss-alignment of the equipment or lights.

The five-choice serial reaction time box was not designed so that mice could meet the criteria for learning. We are currently redesigning the apparatus to eliminate this problem and this will allow us to complete the experiment in less than half the time required in published papers.

The automated recording of freezing behavior in cued and context conditioning resulted in errors in setting the baseline and in recording freezing as discussed by Marchand et al. [8].

The discrepancies between the automated equipment and the observers were due to (1) different operational definitions of the behaviors, (2) equipment hardware and software problems, and (3) improper adjustment of the equipment. We have corrected the hardware, software and equipment adjustment problems but the problem of operational definitions remains an issue.

One of the problems with behavioral phenotyping is that researchers often receive no training in behavior analysis [1]. With the pressure to test more and more animals faster and faster, automated equipment will be used more often, but, while this equipment increases the speed of testing and compares favourably with trained observers on latency measures, the error rate of automated equipment is often unknown. Because test apparatus is not standardized, new apparatus is often not tested parametrically for reliability and validity, and if experimenters are not trained observers of behavior, the equipment errors go undetected. In experiments using automated equipment, we recommend that exact details be given in the methods section of papers, behaviors be videotaped and that the results from automated equipment be

verified by trained observers and that error rates for such equipment be calculated.

References

1. Blizard, D.A., Takahashi, A., Galsworthy, M.J., Martin, B., Koide, T. (2007). Test standardization in behavioural neuroscience: a response to Stanford. *Journal of Psychopharmacology* **21**, 136-139.
2. Brown, R.E. (2007). Behavioural phenotyping of transgenic mice. Canadian *Journal of Experimental Psychology* **61**, 328-344.
3. Crowley, J.J., Jones, M.D., O'Leary, O.F., Lucki, I. (2004). Automated tests for measuring the effects of antidepressants in mice. *Pharmacology, Biochemistry and Behavior* **78**, 269-274.
4. Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L., Mogil, J. S. (2002). Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. *Neuroscience and Biobehavioral Reviews* **26**, 907-923.
5. Fitch, T., Adams, B., Chaney, S., Gerlai, R. (2002). Force transducer-based movement detection in fear conditioning in mice: a comparative analysis. *Hippocampus* **12**, 4 -17.
6. Graziano, A., Petrosini, L., Bartoletti, A. (2003). Automatic recognition of explorative strategies in the Morris water maze. *Journal of Neuroscience Methods* **130**, 33-44.
7. Lehner, P.N. (1996). *Handbook of Ethological Methods*. 2nd ed. Cambridge: Cambridge University Press.
8. Marchand, A.R., Luck, D., DiScala, G. (2003). Evaluation of an improved automated analysis of freezing behaviour in rats and its use in trace fear conditioning. *Journal of Neuroscience Methods* **126**, 145-153.
9. Nadler, J.J., Moy, S.S., Dold, G., Trang, D., et al. (2004). Automated apparatus for quantitation of social approach behaviors in mice. *Genes, Brain and Behavior* **3**, 303-314.
10. Tecott, L. H., Nestler, E. J. (2004). Neurobehavioral assessment in the information age. *Nature Neuroscience* **7**, 462-466.
11. Wahlsten, D., Metten, P., Phillips, T. J., et al.. (2003). Different data from different labs: Lessons from studies of gene-environment interaction. *Journal of Neurbioogy* **54**, 283-311.