# In search of a better mouse test

By: Douglas Wahlsten, Nathan R. Rustay, Pamela Metten, and John C. Crabbe

Wahlsten D, Rustay N, Metten P, Crabbe JC. (2003) In search of a better mouse test. *Trends in Neuroscience*, 26: 132-136.



## Abstract:

To elucidate pathways from specific genes to complex behaviors, assays of mouse behavior need to be valid, reliable and replicable across laboratories. Behavioral assays are proving to be as complex as the intricate cellular and molecular pathways that are the main interest of many mouse users. There is no perfect behavioral test, but we propose some aphorisms to stimulate discussion that is necessary for continued progress in task development. For maximal utility, a behavioral test should yield valid data for most of the commonly used inbred mouse strains. Tests of simple, ubiquitous behaviors usually yield meaningful data for most mice, especially when based on automated scoring or on simple physical measures that are likely to be replicable across laboratories. Extreme test scores resulting from non-performance on a task can inflate the apparent reliability of a test, and devious adaptations to a task can undermine its validity. The optimal apparatus configuration for certain genetic or pharmacological analyses might depend on the particular laboratory environment. Despite our best efforts, the mice will continue to win some innings.

## Article:

Many genetic mutations in mice alter the nervous system and thereby change behavior. Consequently, mouse behavior is a phenotype of considerable interest in neuroscience, and comprehensive assessment of a mutation must involve assays of behavior. Although numerous tests of behavior are available [1], we believe that some of the popular tests can be made more reliable, valid and replicable across laboratories. Here, we discuss criteria for building better mouse tests. Behavioral testing should be viewed as a work in progress – an enterprise made more challenging by the daunting complexity of behavior and its sensitivity to factors that are quite subtle [2].

## Diversity, the raw material for test evolution

The search for better tests is made a little easier because there is no standard implementation of any single test that is deeply entrenched. Instead, most tests are done in a manner that is unique to each laboratory [3]. For example, on the submerged-platform water-escape task, details of the apparatus and procedures vary widely (Fig. 1). There is considerable evidence that parametric differences in the details of many tasks are important for the outcome of genetic experiments [4–6]. Minor changes in a task can sometimes yield large benefits, as exemplified by the rescue of maze learning in staggerer mutant mice when walls are added to the arms of an elevated maze [7] and by the improved care of pups observed when staggerer mothers and their litters are housed in a cone-shaped cage [8].

Behavioral testing is central to the phenotyping initiatives at the National Institute of Mental Health [9] (see also What's Wrong with My Mouse? at http://www. mymouse.org), the Mouse Phenome Project (MPP) [10] and the Mouse Phenome Database (http://www.jax.org/phenome). Some of the strains proposed by the MPP present special challenges for behavioral testing (Fig. 2). For example, many have retinal degeneration that can alter behavior in tasks [11–13]; others have age-related hearing loss [14], and wild-derived strains are very difficult to handle in behavioral studies [15]. The wide range of strains also includes some real gems, such as the BTBR T+ *tf*/*tf* strain, which shows extraordinarily good rotarod performance [16,17], even though a survey

of 21 inbred mouse strains carried out in two laboratories reveals that BTBR T+ tf/tf have a severely reduced hippocampal commissure and lack the corpus callosum [18].

## Inbred strains and mutants: one world

The need to gather behavioral data on a wide range of inbred strains brings into focus the shortcomings of some tests. To have maximal utility, a good behavioral test should yield valid data for most of the commonly used inbred strains. Of course, much contemporary research in neuroscience involves single-gene mutations rather than inbred strains, but the conclusion applies equally to mutants. It is widely appreciated that some strains do not perform well on certain tests [19,20], and the wisdom of choosing a 'wild-type' strain that does well on a task as the genetic background of a mutation is generally understood (if not practiced). A less obvious problem can arise when a mutation placed on the high-performance background causes a deficit in performance for the same reason that invalidates the task for certain inbred strains. If the test works well for only a few strains, then perhaps the task itself needs to be improved or even replaced.

When evaluating and refining tests, there are advantages to working in collaboration with several laboratories, with all researchers agreeing to adopt identical physical apparatus and test protocols that yield valid results for most mice [2,21]. The iterative processes of consultation and negotiation identify procedures that are often carried out without any good justification. An alternative is to adopt the methods of one laboratory as the 'gold standard' – an approach that risks substituting ego for evidence in the choice of task parameters.

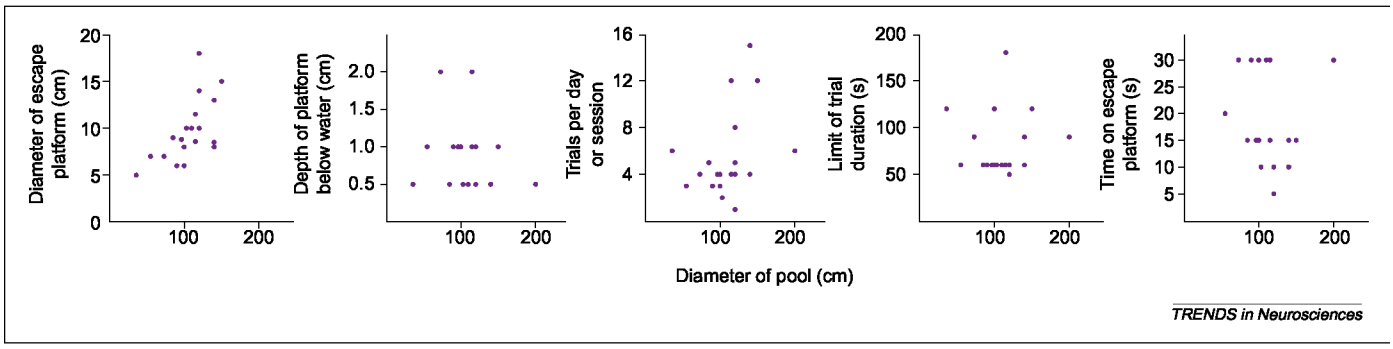


**Fig. 1.** Scatter plots of five task parameters versus the diameter of the water tank in 22 studies of mice on the submerged-platform water-escape task [31–52]. The studies were published from 1996 to 2001 and were identified through a search of Current Contents (http://www.isinet.com/isi/products/cc/). Not all points are shown for all variables because some studies did not report all six parameters.

## The simplest case: exploratory activity

Because deficits are often expected in a study of mutations, a crucial consideration for any test is the meaning of low or zero test scores. It is important that mice do not fail for a reason that is unrelated to what the test purports to measure. We assert that tests of simple, ubiquitously performed behaviors usually yield meaningful data for most mice. For example, the 129S1/SvImJ strain exhibits surprisingly low exploration on the second test trial in an open field or Y-maze (Fig. 3), often sitting quietly in one arm of a Y-maze for an entire trial. Call it boredom if you like, but this pattern clearly requires good long-term memory. But exploration itself is not a test of memory: several other inbred strains (e.g. BALB/cByJ and DBA/2J) show high levels of initial exploration day after day, even though it is known from other studies that these mice have a reasonably good memory for places. Exploration of a simple environment involves gross motor activity, and distance traveled is a good measure of exploratory behavior. A score of nearly zero centimeters traveled does not tell us whether the mouse was sleeping, grooming or paralyzed by anxiety, but it does tell us that the mouse was not exploring its environment.

A single measure such as distance traveled cannot encapsulate everything that a mouse does; additional measures of where the mouse goes are needed to distinguish between repetitive circling and exploration. It is unlikely that any reasonably complex behavior can be represented adequately by a single measure. The variety of behaviors documented and the methods for analyzing patterns of movement can be enhanced [22], and absolute scores will certainly differ with time of day, level of illumination and many other parameters.

Sophisticated descriptors of locomotion can detect consistent patterns of strain differences across laboratories [23], yet distance traveled remains a highly reliable and valid indicator of exploration.

### Human versus machine: which to trust?

Another meaningful zero score occurs in a test of hunger, in which mice are observed for a short time in the home cage eating their usual laboratory chow that is presented in a familiar glass dish. The amount of food eaten increases steadily with the duration of deprivation. But there are two ways in which to measure eating: the actual mass of food consumed and the time spent eating, as judged by a trained observer. Within a single laboratory these two measures correlate strongly, but when we made the measurements simultaneously in Edmonton and Portland it became apparent that our technicians used different criteria to record time spent eating the same mass of food, despite detailed written instructions.

We conclude that automated scoring by machine or reliance on simple physical measures is more likely to yield an index of behavior that is replicable across laboratories. The impact of the experimenter on the data is likely to be greatest when the task involves extensive or expert handling of the [6]. When activity or hunger levels are lower in one laboratory than in another, this can be interpreted only if identical measures are being used. Ethologically meaningful measures of behavior are important, but implementing such measures with automation is desirable [13,22].

### The dilemma of uncooperative mice

Some tests are effectively undermined by non-performance. For 129S1/ SvImJ mice (Fig. 3), the index of alternation between arms of a Y-maze on the second day often cannot be computed owing to a lack of exploration. A similar problem is encountered with the Barnes maze, a test of spatial memory in which mice must locate an open hole among 12 or more holes along the periphery of a large disk to escape from bright light. Many mice ‘freeze’ in the center of the disk and never escape at all. For these non-performing mice, the Barnes maze yields no valid data. Paradoxically, including such mice in the final analysis can make the test look deceptively good. Test–retest reliability (the correlation of performance on different test occasions) tends to be high when there are extreme scores by some individuals on both occasions. For use of the Barnes maze in Edmonton, reliability is quite high ($r = 0.7$) when non-performing mice are included in the analysis but less so ($r = 0.5$) when those that fail to escape within 60 s are eliminated. Thus, caution is warranted when interpreting data from some tasks, because extreme test scores resulting from non-performance on a task can inflate the apparent reliability of a test at the expense of its validity.

The elevated plus maze – a test of mouse anxiety in which anxious mice avoid open arms and ambulate in arms with high walls – is undermined when the mice remain at the center hub or are inactive for most of a trial. Exploration of open arms correlates with overall levels of activity [24]. This issue could be addressed by removing inactive mice from the analysis [21], but this might bias the results of a genetic experiment. Rather than eliminating data, perhaps experimenters should alter the task itself. In Portland, mice show a much lower level of exploration of open arms on the plus maze than do mice in Edmonton on identical mazes [2]. A low baseline activity on the plus maze has been reported in several other laboratories, precluding its use in studies of drugs or mutations that might be anxiogenic. This problem might be ameliorated by using a slightly higher rim on the open arms of the maze, which increases the amount of open-arm exploration in Portland to the level seen previously in Edmonton (Fig. 4). In another test of anxiety, the mirrored chamber, redesign of apparatus and procedures allows the detection of increased anxiety (C. Kliethermes et al., unpublished).

Thus, we reluctantly conclude that, for tests that are especially sensitive to laboratory conditions, the optimal apparatus configuration for certain genetic or pharmacological experiments could depend on the particular laboratory environment. Equating apparatus and proto-cols across laboratories is highly desirable, but differences in apparatus sometimes might be required to yield equivalent behaviors in different laboratories.

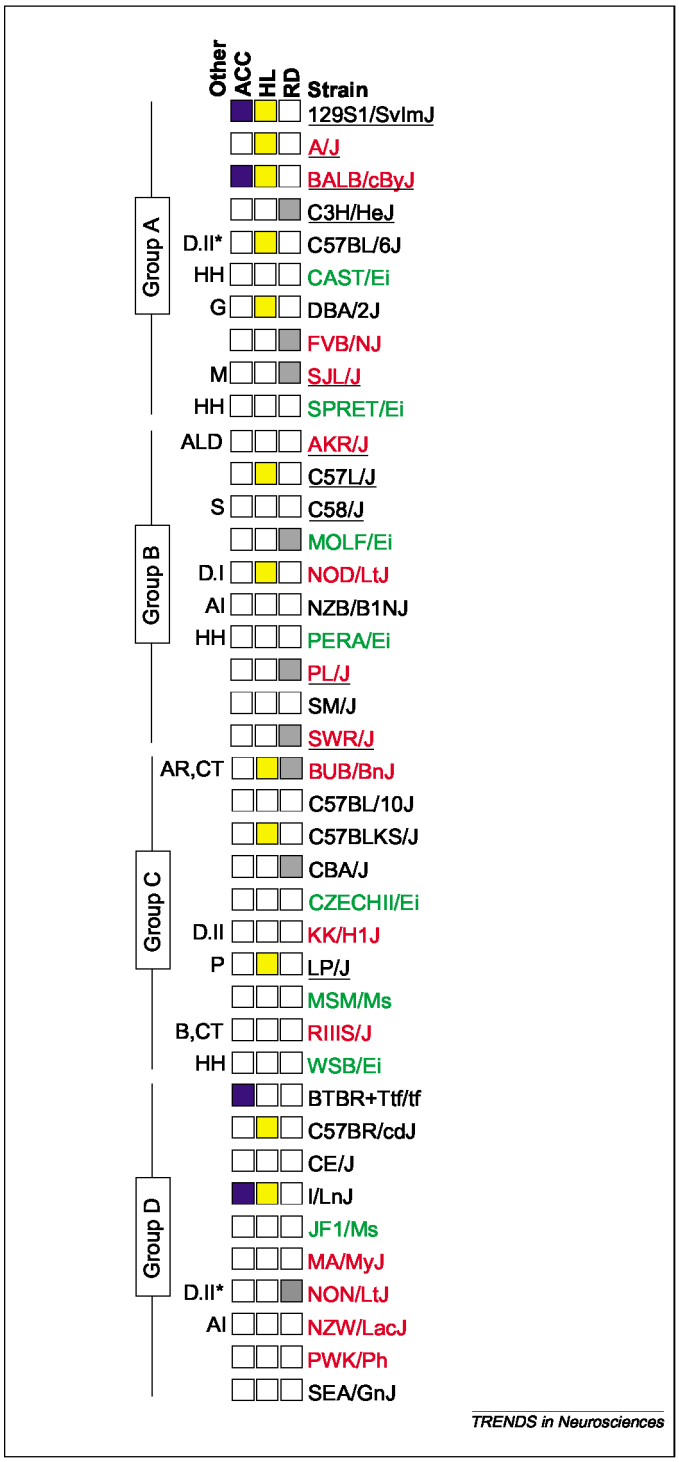


**Fig. 2.** Some salient features of strains selected for the Mouse Phenome Project in four priority groups, with Group A having the highest priority for phenotyping. Strains shown in red type are albino and can be expected to have mild vision defects [29,30]. Strains in green type are wild-derived and often are hard to handle in behavioral studies. Strains underlined often have cancerous tumors present in older mice. Major defects present in four or more strains include absent corpus callosum (ACC; dark blue), hearing loss (HL; yellow) and retinal degeneration (RD; gray (http://www.jax.org). Other defects or unusual features include autoimmune abnormalities (AI), adrenal lipid depletion (ALD), arthritis (AR), blood-clotting deficiency (von Willebrand disease model) (B), cataracts (CT), diabetes Type I or II (D.I or D.II, respectively; asterisk indicates diabetes seen in mice on a high-fat diet), glaucoma (G), the mice being hard to handle [15] (HH), myopathy and muscle weakness (M), piebald spotting (Hirschprung disease model) (P), and sinking in water-escape tests (S). Data were compiled from the Jackson Laboratory website on mouse models (http://jaxmice.jax.org/jaxmicedb/html/models.shtml) and observations from our laboratories. Not all of the 40 strains have been assessed for all phenotypes.

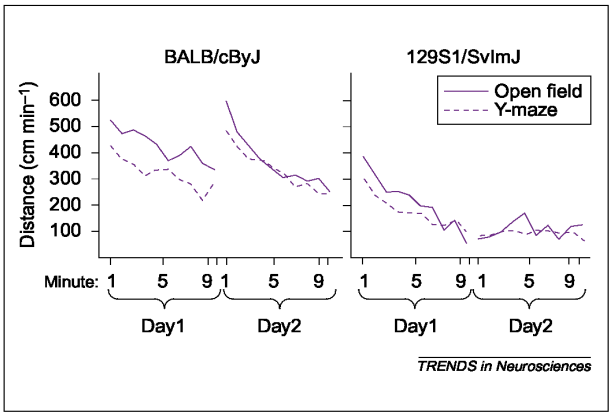


**Fig. 3.** Distance traveled per minute in a box of 50 cm by 50 cm, or in a Y-maze with three identical arms, by two strains of mice during a 10-min test on two successive days in Edmonton. Tracking was carried out using identical VideoScan™ systems (AccuScan Inc.) for both sets of apparatus. The 129S1/SvImJ strain is unique among those tested so far in that its activity does not recover at the start of the trial on the next day.

## The challenge of devious adaptations

Equally vexing are devious adaptations to complex tasks. Mice trained to avoid electric shock in a two-way shuttle box often resort to clever ruses to avoid shock without running, including widely splayed legs to bridge grid bars of the same voltage safely, standing on one bar and leaning nonchalantly against a plastic wall, and finding a safe zone between photocell beams. Even the simple rotarod test of motor coordination can be circumvented by our furry subjects. Many mice flatten themselves and hold tight to the rotating rod to ride passively instead of running on top of it, thereby compromising the validity of the test. It turns out that a larger diameter rod (6.5 cm; the 'rat' size supplied by some manufacturers) covered with number-320 grit emery paper almost completely eliminates passive rotation by mice and yields excellent data for evaluating the effects of ethanol on the coordination of many inbred strains [16]. Because the rotarod surface gradually wears with use, like an automobile tire, regular renewal of the surface gives stable results over long periods of time – a feature that can be important for studies of aging.

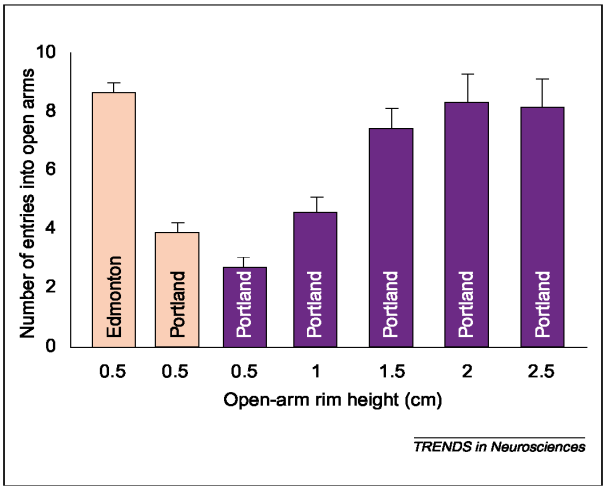


**Fig. 4.** Number of entries into the open arms of identical elevated plus mazes in two different laboratories in a 5-min test. Each plus maze has four arms that are 5 cm wide by 30 cm long, two of which are 'closed' (with 15-cm-high clear plastic walls) and two of which are 'open' (with low rims rather than high walls). Data collected in the Edmonton and Portland laboratories in 1998 (pink bars) [2] have been averaged across the same eight strains. Additional data from testing in Portland with different open-arm rim heights (purple bars) have been averaged across several other genotypes, including Withdrawal Seizure-Prone and Seizure-Resistant replicated selection lines developed by J. Crabbe and a 129 Sv/ter strain developed by R. Hen. In the more recent data, the open-arm wall height is significantly ($P < 0.01$) related to the amount of open-arm exploration but not to the total entries into all four arms. An open-arm wall height of ≥1.5 cm in Portland yields about the same amount of open-arm exploration as seen in Edmonton with walls of 0.5 cm.

## Sink, swim or float: those are the options

The advantages and disadvantages of using escape from water to motivate learning in mice are well known [25,26]. Neuroscientists have been warned that many strains perform poorly on the submerged-platform water-escape test task [19], which is better suited to rats than to mice [27], yet it is used widely for the study of memory in mice. The A/J strain is an implacable wall-hugger in the circular water tank [21]. Most of the variance among different mouse strains and mutants is attributable to wall-hugging and floating, and not to spatial memory [28]. The way in which the test is done is known to be very important [5], so perhaps the task can be altered to rescue poor performance. Adding clear plastic walls to create four wide arms, while keeping the escape platform and extra-maze stimuli the same as in the open version of the tank, might allow A/J mice to remain close to a wall to reach the platform. However, this is a nice idea that proves to be wrong for this mouse strain. When there is a submerged platform in every arm of the water maze, so that there is an easy way out whichever arm is chosen, almost all mice learn to swim very quickly; by contrast, when they are required to discriminate among options to find the platform, many mice give up and float like little corks (Fig. 5).

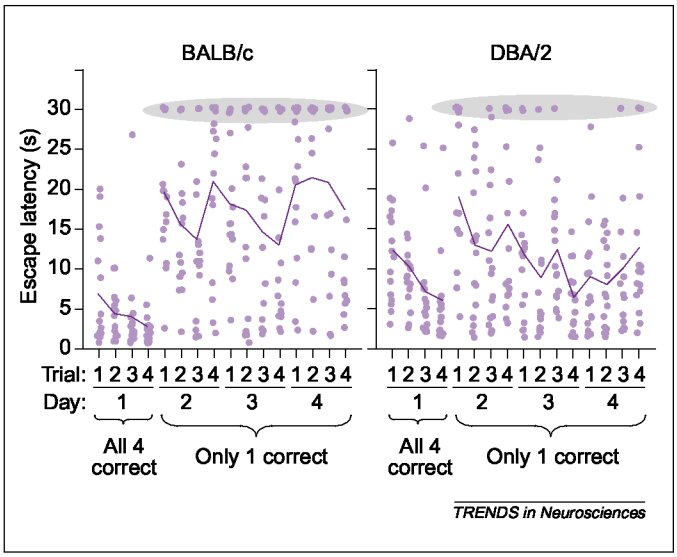


**Fig. 5.** Escape latencies in a 70-cm-diameter water maze with submerged platforms, in which the maze is divided by clear plastic walls into four arms with a width of 10 cm. Points for individual mice are jittered slightly to reveal overlapping scores. Mice (16 per strain) underwent four trials per day. On the first day, every arm had an escape platform, and all mice learned quickly to swim to the end of an arm. On the next three days, however, only one arm has an escape platform and mice must use room cues to locate it. Escape latencies initially increase for both strains on day 2 but, whereas DBA/2 mice gradually learn the platform location, many BALB/c mice cease swimming and instead float. The proportion of mice reaching the 30-s trial limit (gray zone) decreases rapidly for the DBA/2 mice but not the BALB/c. Mice were obtained from Charles River Co.

The search continues for satisfactory results for most strains on water-escape learning, but our initial enthusiasm for the quest has been dampened considerably. Perhaps the submerged-platform water-escape test should be replaced as the task of first choice for evaluating memory, rather than rejecting all mouse strains that persist in floating, hugging the walls or slowly sinking beneath the waves (the C58/J strain).

## An obvious problem with no easy solution

Further work needs to be done on several behavioral tests to ensure that valid data can be obtained for most of the commonly used inbred mouse strains. If low scores are not interpretable for some inbred strains because of non-performance or devious adaptations, then it is possible that a knockout strain might perform poorly for the same reasons. If one wishes to explore the genetics of memory, for example, it is essential that low scores on a task be valid indicators of poor memory. This point is simple, even obvious, but the solution to the challenges posed by mice of different genotypes is neither simple nor obvious for complex behaviors.

## References

1 Crawley, J.N. (2000) What's Wrong With My Mouse? Behavioral Phenotyping of Transgenic and Knockout Mice, Wiley-Liss

2 Crabbe, J.C. et al. (1999) Genetics of mouse behavior: interactions with laboratory environment. Science 284, 1670–1672
3 Wahlsten, D. (2001) Standardizing tests of mouse behavior: reasons, recommendations, and reality. Physiol. Behav. 73, 695–704
4 Boehm, S.L. II et al. (2000) Sensitivity to ethanol-induced motor in coordination in 5-$HT_{1B}$ receptor null mutant mice is task-dependent: implications for behavioral assessment in genetically altered mice. Behav. Neurosci. 114, 401–409
5 Gerlai, R. (2001) Behavioral tests of hippocampal function: simple paradigms, complex problems. Behav. Brain Res. 125, 269–277
6 Chesler, E.J. et al. (2002) Influences of laboratory environment on behavior. Nat. Neurosci. 5, 1101–1102
7 Goldowitz, D. and Koch, J. (1986) Performance of normal and neurological mutant mice on radial arm maze and active avoidance tasks. Behav. Neural Biol. 46, 216–226
8 Guastavino, J-M. (1984) Environmental features determining successful rearing in the mutant mouse staggerer. Physiol. Behav. 32, 225–228
9 Moldin, S.O. et al. (200 1) Trans-NIH neuroscience initiatives on mouse phenotyping and mutagenesis. Mamm. Genome 12, 575–581
10 Paigen, K. and Eppig, J.T. (2000) A mouse phenome project. Mamm. Genome 11, 715–717
11 Drager, U.C. and Hubel, D.H. (1978) Studies of visual function and its decay in mice with hereditary retinal degeneration. J. Comp. Neurol. 180,85–114
12 Cook, M.N. et al. (200 1) Anxiety-related behaviors in the elevated zero-maze are affected by genetic factors and retinal degeneration. Behav. Neurosci. 115, 468–476
13 Gerlai, R. (2002) Phenomics: fiction or the future? Trends Neurosci. 25, 506–509
14 Zheng, Q.Y. et al. (1999) Assessment of hearing in 80 inbred strains of mice by ABR threshold analyses. Hear. Res. 130, 94–107
15 Wahlsten, D. et al. A rating scale for wildness and ease of handling laboratory mice: results for 21 inbred strains tested in two laboratories. Genes Brain Behav. (in press)
16 Rustay, N.R. et al. Influence of task parameters on rotarod performance and sensitivity to ethanol in mice. Behav. Brain Res. (in press)
17 Rustay, N.R. et al. Assessment of genetic susceptibility to ethanol intoxication in mice. Proc. Natl. Acad. Sci. U. S. A. (in press)
18 Wahlsten, D. et al. Survey of 21 inbred mouse strains in two laboratories reveals that BTBR T/ + tf/tf has severely reduced hippocampal commissure and absent corpus callosum. Brain Res. (in press)
19 Crawley, J.N. et al. (1997) Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. Psychopharmacology (Berl.) 132, 107–124
20 Tarantino, L.M. et al. (2000) Behavior and mutatgenesis screens: the importance of baseline analysis of inbred strains. Mamm. Genome 11, 555–564
21 Wahlsten, D. et al. (2003) Different data from different labs: lessons from studies of gene-environment interaction. J. Neurobiol. 54, 283–311
22 Drai, D. and Golani, I. (2001) SEE: a tool for the visualization and analysis of rodent exploratory behavior. Neurosci. Biobehav. Rev. 25, 409–426
23 Kafkafi, N. et al. SEE analysis of open-field behavior discriminates C57BL/6J and DBA/2J mouse inbred strains across laboratories and protocol conditions. Behav. Neurosci. (in press)
24 Weiss, S.M. et al. (1998) Utility of ethological analysis to overcome locomotor confounds in elevated maze models of anxiety. Neurosci. Biobehav. Rev. 23, 265 – 271
25 Festing, M. (1973) Water escape learning in mice. I. Strain differences and biometrical considerations. Behav. Genet. 3, 13–24
26 Festing, M. (1973) Water escape learning in mice. II. Replicated selection for increased learning speed. Behav. Genet. 3, 25–36
27 Whishaw, I.Q. and Tomie, J-A. (1996) Of mice and mazes: similarities between mice and rats on dry land but not water mazes. Physiol. Behav.60,1191–1197
28 Wolfer, D.P. et al. (1998) Spatial memory and learning in transgenic mice: fact or artifact? News Physiol. Sci. 13, 118–122

29 Guillery, R.W. (1974) Visual pathways in albinos. Sci. Am. 230,44–54
30 Rice, D.S. et al. (1995) Genetic control of retinal projections in inbred strains of albino mice. J. Comp. Neurol. 354, 459–469
31 Chapillon, P. (1999) Very brief exposure to visual distal cues is sufficient for young mice to navigate in the Morris water maze. Behav. Proc. 46, 15–24
32 Chapillon, P. and Debouzie, A. (2000) BALB/c mice are not so bad in the Morris water maze. Behav. Brain Res. 117, 115–118
33 Czech, D.A. et al. (2000) Chronic propranolol induces deficits in retention but not acquisition performance in the water maze in mice. Neurobiol. Learn. Memory 74, 17–26
34 D'Hooge, R. et al. (1997) Mildly impaired water maze performance in male Fmr1 knockout mice. Neuroscience 76, 367 –376
35 Dalm, S. et al. (2000) Quantification of swim patterns in the Morris water maze. Behav. Res. Meth. Instr. Comp. 32, 134–139
36 Francis, D.D. et al. (1995) Stress-induced disturbances in Morris water-maze performance: interstrain variability. Physiol. Behav. 58, 57–65
37 Frick, K.M. et al. (2000) Mice are not little rats: species differences in a one-day water maze task. Neuroreport 11, 3461–3465
38 Frisch, C. et al. (2000) Superior water maze performance and increase in fear-related behavior in the endothelial nitric oxide synthase-deficient mouse together with monoamine changes in cerebellum and ventral striatum. J. Neurosci. 20, 6694–6700
39 Hengemihle, J.M. et al. (1996) Chronic treatment with human recombinant erythropoietin increases hematocrit and improves water maze performance in mice. Physiol. Behav. 59, 153–156
40 Holcomb, L.A. et al. (1999) Behavioral changes in transgenic mice expressing both amyloid precursor protein and presenilin-1 mutations: Lack of association with amyloid deposits. Behav. Genet. 29,177–185
41 Holschneider, D.P. et al. (1999) Lack of protection of monoamine oxidase B-Deficient mice from age-related spatial learning deficits in the Morris water maze. Life Sci. 65, 1757–1763
42 Lee, B. et al. (2000) LP-BM5 infection impairs spatial working memory in C57BL/6 mice in the Morris water maze. Brain Res. 856, 129–134
43 Logue, S.F. et al. (1997) Hippocampal lesions cause learning deficits in inbred mice in the Morris water maze and conditioned-fear task. Behav. Neurosci. 111, 104–113
44 Malleret, G. et al. (1999) 5-HT1B receptor knock-out mice exhibit increased exploratory activity and enhanced spatial memory performance in the Morris water maze. J. Neurosci. 19, 6157–6168
45 Oitzl, M.S. et al. (1997) Severe learning deficits in apolipoprotein E-knockout mice in a water maze task. Brain Res. 752, 189–196
46 Rissanen, A. et al. (1999) In mice tonic estrogen replacement therapy improves non-spatial and spatial memory in a water maze task. NeuroReport 10, 1369–1372
47 Roder, J.K. et al. (1996) Memory and the effect of cold shock in the water maze in S100P transgenic mice. Physiol. Behav. 60, 611–615
48 Vicens, P. et al. (1999) Previous training in the water maze: differential effects in NMRI and C57BL mice. Physiol. Behav. 67, 197–203
49 Watanabe, C. and Satoh, H. (1995) Effects of prolonged selenium deficiency on open field behavior and Morris water maze performance in mice. Pharmacol. Biochem. Behav. 51, 747–752
50 Wilson, I.A. et al. (1999) Estrogen and NMDA receptor antagonism: effects upon reference and working memory. Eur. J. Pharmacol. 381, 93–99
51 Yukawa, K. et al. (2000) Impaired water-maze performance in mice lacking transcription factor CCAAT/Enhancer-binding protein Neurosci. Res. Comm. 26,59–67
52 Zaharia, M.D. et al. (1996) The effects of early postnatal stimulation on Morris water-maze acquisition in adult mice: genetic and maternal factors. Psychopharmacology 128, 227–239