Behavioral Neuroscience
2003, Vol. 117, No. 3, 464–477

Copyright 2003 by the American Psychological Association, Inc.
0735-7044/03/$12.00 DOI: 10.1037/0735-7044.117.3.464

SEE Locomotor Behavior Test Discriminates C57BL/6J and DBA/2J Mouse Inbred Strains Across Laboratories and Protocol Conditions

Neri Kafkafi
National Institute on Drug Abuse and University of Maryland

Dina Lipkind and Yoav Benjamini
Tel Aviv University

Cheryl L. Mayo and Gregory I. Elmer
University of Maryland

Ilan Golani
Tel Aviv University

Conventional tests of behavioral phenotyping frequently have difficulties differentiating certain genotypes and replicating these differences across laboratories and protocol conditions. This study explores the hypothesis that automated tests can be designed to quantify ethologically relevant behavior patterns that more readily characterize heritable and replicable phenotypes. It used SEE (Strategy for the Exploration of Exploration) to phenotype the locomotor behavior of the C57BL/6 and DBA/2 mouse inbred strains across 3 laboratories. The 2 genotypes differed in 15 different measures of behavior, none of which had a significant genotype–laboratory interaction. Within the same laboratory, most of these differences were replicated in additional experiments despite the test photoperiod phase being changed and saline being injected. Results suggest that well-designed tests may considerably enhance replicability across laboratories.

Neurobehavioral genetics depends critically on the accuracy and consistency of behavioral measurements. The characterization of specific behaviors, sometime referred to as *behavioral phenotyping*, is essential for eventually associating them with particular gene loci. The need for behavioral phenotyping has resulted in the design of behavioral and physiological test batteries for mice (Crawley, 2000; Crawley et al., 1997; Rogers et al., 1999). Considerable effort has been made to automate these tests to increase the throughput needed for testing large numbers of animals and to avoid the effect of subjective human judgment. A locomotor behavior test in an unfamiliar arena is included in most test batteries and can be conducted efficiently with standard, commercially available photobeam systems.

Neri Kafkafi, Behavioral Neuroscience Branch, National Institute on Drug Abuse/Intramural Research Program, Baltimore, Maryland, and Department of Psychiatry, Maryland Psychiatric Research Center, School of Medicine, University of Maryland; Dina Lipkind and Ilan Golani, Department of Zoology, Tel Aviv University, Tel Aviv, Israel; Yoav Benjamini, Department of Statistics, Tel Aviv University; Cheryl L. Mayo and Gregory I. Elmer, Department of Psychiatry, Maryland Psychiatric Research Center, School of Medicine, University of Maryland.

This research was supported by National Institutes of Health Grant 1 R01 NS4023.

The SEE software is available from the authors on request. It requires, however, the *Mathematica* programming environment by Wolfram Research, Inc. A simpler, stand-alone program for automatic smoothing, segmentation, and calculation of endpoints is also available. We thank Noldus Information Technology for the use of their EthoVision system at Tel Aviv University.

Correspondence concerning this article should be addressed to Neri Kafkafi, Department of Psychiatry, Maryland Psychiatric Research Center, School of Medicine, University of Maryland, P.O. Box 21247, Baltimore, Maryland 21224. E-mail: nkafkafi@mprc.umaryland.edu

C57BL/6 (B6) and DBA/2 (D2) are two of the most commonly used inbred strains of laboratory mice. Consequently, many behavioral differences between them have been reported in the literature (for a review, see Crawley et al., 1997). Because the locomotor behavior test is one of the most common tests, the locomotor behavior of these two genotypes has been reported in many studies. Common view holds that B6 is a high-activity strain, whereas the D2 is an intermediate-activity strain (Cabib, 2002; Crawley et al., 1997). This view, however, seems to represent only a broad average over studies, with some that reported B6 to be significantly more active (e.g., Bolivar, Caldarone, Reilly, & Flaherty, 2000, for females; Elmer, Gorelick, Goldberg, & Rothman, 1996; Gorris & van Abeelen, 1981; Hatcher et al., 2001; Logue, Owen, Rasmussen, & Wehner, 1997), many that did not detect a significant difference in activity (e.g., Bolivar et al., 2000, for males; Cabib & Bonaventura, 1997; Jones, Reed, Radcliffe, & Erwein, 1993; Rocha et al., 1998; Tirelli & Witkin, 1994; Tolliver & Carney, 1995; Womer, Jones, & Erwin, 1994), and at least one (Rogers at al., 1999) that found D2 mice to have significantly higher activity. It should be noted that these studies used different arena sizes (longer dimension of 40–60 cm), arena shapes (square, rectangular, or circular), light conditions (fully illuminated to complete darkness), and tracking techniques (photocells, photobeam, and video). There was no apparent relationship, however, between any of these factors and the ability to differentiate the two strains.

The problem of replicating behavioral results in other laboratories is particular neither to these two genotypes nor to the locomotor behavior test. Consequently, Crabbe, Wahlsten, and Dudek (1999a) conducted a pioneering study in which they compared eight genotypes in a battery of several standard behavioral tests across three laboratories. Despite the rigorous standardization of tests and housing protocols, they reported many significant lab and

Lab × Genotype interaction effects. One of their conclusions was that genotype differences found in a single laboratory might prove to be idiosyncratic to this laboratory. This conclusion might be interpreted as pointing to a major hindrance in current behavioral genetics research. As such, the importance of replicable phenotypes drives prominent database organizers to require submitted data to be validated in at least two different laboratories (Mouse Phenome Database; Paigen & Eppig, 2000).

Crabbe et al. (1999a) included the B6 and D2 as part of the eight genotypes in their study but did not examine the differences between these two genotypes separately. In another pioneering step, however, they published all the raw results of their study in a Web site, in a convenient format for downloading and analyzing. We have taken advantage of this feature to estimate the discriminative power and replicability of the new locomotor behavior method described in this study relative to a commonly used photobeam system for these two genotypes. As we suggest in the Discussion, both the multilab studies and the publication of the raw results on the Web constitute a fruitful approach to behavioral phenotyping. An examination of the results from this Web site shows that most measures of the locomotor behavior test either did not detect significant differences between B6 and D2 or (in agreement with the general conclusion of Crabbe et al.'s report) revealed differences that were not consistent across the three laboratories.

The main remedy advocated so far for the lack of replicability is more careful standardization of test protocol, handling procedures, and laboratory environment (van der Staay & Steckler, 2002; Wahlsten, 2001; Wahlsten, Crabbe, & Dudek, 2001; but see Wurbel, 2000, 2002). This remedy, however, is expected to require a considerable effort (Wahlsten, 2001), as the level of standardization in Crabbe et al. (1999a) was already much higher than is currently practiced in the field. We suggest a complementary approach: Design improved standard tests that can capture ethologically relevant behavior patterns more precisely. Such tests may be more resistant to the laboratory environment and to small changes in protocol details. Locomotor behavior constitutes a good test case for this approach. Current locomotor behavior tests for mice, usually using photobeam systems, are conducted in small cages of 40–60 cm width. They typically use simple measures such as the distance traveled by the animal and the time spent in the center of the arena. These measures are cumulative and general, reflecting a common view that locomotor behavior is largely stochastic in nature and can be quantified mainly by some measure of general activity (but see Paulus & Geyer, 1993, for a different viewpoint). In recent years, however, ethologically oriented studies in rats (Drai, Benjamini, & Golani, 2000; Eilam & Golani, 1989; Eilam, Golani, & Szechtman, 1989; Golani, Benjamini, & Eilam, 1993; Kafkafi, Mayo, Drai, Golani, & Elmer, 2001; Tchernichovski, Benjamini, & Golani, 1996, 1998) and, more recently, in mice (Benjamini, Drai, Elmer, Golani, & Kafkafi, 2001; Drai, Kafkafi, Benjamini, Elmer, & Golani, 2001; Kafkafi et al., in press) found that locomotor behavior is highly structured and consists of typical behavior patterns. Once these patterns were isolated, they were found useful in psychopharmacological and psychobiological studies (Cools, Ellenbroek, Gingras, Engbersen, & Heeren, 1997; Gingras & Cools, 1997; Szechtman, Culver, & Eilam, 1999; Whishaw, Cassel, Majchrzak, Cassel, & Will, 1994; Wallace, Hines, & Whishaw, 2002; Whishaw, Hines, & Wallace, 2001). On the basis of these patterns, a Strategy for the Exploration of Exploration (SEE), complemented by respective software, was recently developed for the visualization and analysis of locomotor behavior data measured automatically by video tracking (Drai & Golani, 2001), and it was proposed as a tool for behavioral phenotyping (Benjamini et al., 2001; Drai et al., 2001).

The SEE locomotor behavior test consists of both the SEE software and several testing procedures that were developed with it to enhance the accuracy, reliability, and relevance of measurement. The SEE test has several properties that are suggested by the results of previous and present phenotyping projects to be important for genotype discrimination:

1. Large (2.50-m diameter) circular arena increases the area by 25 to 80 times that of common photobeam systems. Combined with a slightly better spatial resolution because of the use of video tracking, this means that the number of different locations that can be discriminated by the system is increased by a factor of approximately 100. The large arena also enables the animal to generate a much wider range of speeds, a key measure in the analysis. In addition, our results suggest that the open space of the large arena is more intimidating and consequently accentuates differences in wall hugging behavior.

2. A tracking rate of 25 or 30 frames per second is considerably higher than is currently practical with many photobeam systems. Such temporal resolution is important, as a mouse can accelerate, slow down, and even stop and start again more than once during a single second.

3. Robust smoothing algorithms considerably reduce tracking noise and outliers, which are typical of the output of tracking systems of all types. Many endpoints, including the widely used distance traveled, are sensitive to such noise and artifacts.

4. The path of the animal is automatically segmented into discrete behavioral units with proven ethological relevance for rodents: progression segments separated by stops (lingering episodes; Drai et al., 2000; Kafkafi et al., 2001). Most SEE endpoints use simple properties of such segments, such as their length, duration, and maximal speed. Treating the path as a string of discrete, relevant units rather than a continuous series of coordinates allows a more straightforward analysis of complex structures.

5. The SEE language can be used to query, visualize, and quantify complex properties of the behavior in a database including many sessions from many experiments and to easily design new endpoints for better genotype discrimination and replicability (Drai & Golani, 2001; Kafkafi, in press; Kafkafi et al., in press).

6. The issue of multiple comparisons, arising because of the use of many endpoints, is handled by the false discovery rate approach (Benjamini et al., 2001; Benjamini & Hochberg, 1995). This approach is preferable to both the too-restrictive Bonferroni-like criterion and the too-permissive approach of not controlling for multiple comparisons at all.

In this study we provide an initial examination of the ability of the improved locomotor behavior test with SEE analysis to discriminate mouse genotypes in a replicable way, by phenotyping B6 and D2 mice across three laboratories.

Method

The experiments were conducted in three laboratories: the National Institute of Drug Abuse/IRP laboratory in Baltimore (NIDA), the Maryland Psychiatric Research Center (MPRC) of the University of Maryland, and Tel Aviv University (TAU). There were slight differences between the laboratories in arena size (because of room size limitation), tracking rate (because of the use of the European PAL video system in TAU instead of the American NTSC), and spatial resolution (because of camera parameters and height). These differences are summarized in Table 1 (left section). In addition, two other experiments were used to assess the replicability across experiments with different protocol conditions within the same lab (Table 1, right section): Experiment MPRC/Light (MPRC/L) was performed in the MPRC, but with the mice tested during the light phase of their photoperiod instead of the dark phase. Experiment NIDA/Light–Saline (NIDA/LS) was performed in NIDA with the mice tested at their light phase and also injected with saline immediately before introduction into the arena. In addition, we included some results from a sixth experiment, NIDA/CXBK (NIDA/C), which compared B6 with CXBK/ByJ (CXBK; an inbred recombinant strain originating from a cross between B6 and BALB/c) in NIDA. The time period between any two experiments within the same laboratory was at least 3 weeks. Other than the differences given in Table 1, all conditions were equalized as described below. The animals used in this study were maintained in facilities fully accredited by the American Association for the Accreditation of Laboratory Animal Care (MPRC and NIDA) or by National Institutes of Health (NIH) Animal Welfare Assurance Number A5010-01 (TAU). The studies were conducted at all three locations in accordance with the *Guide for Care and Use of Laboratory Animals* provided by the NIH.

Subjects

The mice were 9–14-week-old males from the inbred strains B6, D2, and CXBK, shipped from Jackson Laboratories (B6, D2) or bred at NIDA (CXBK). Group sizes are given in parentheses in Table 1.

Housing

Animals were kept in a 12-hr light–dark cycle, housed 2–4 per cage under standard conditions of 22 °C room temperature and water and food ad libitum. The animals were housed in their room for at least 2 weeks before the experiment.

Tracking Protocol

Each animal was brought from its housing room, introduced immediately into the arena, and returned after the end of the 30-min session. The arena was a large (210–250-cm diameter), circular area with a nonporous gray floor and a 50-cm-high, primer gray-painted, continuous wall. The gray paint was especially chosen to provide a high-contrast background, enabling video tracking of black, white, brown, and agouti-color mice without the need to dye or mark them. Several landmarks of various shapes and sizes were attached in different locations to the arena wall and to the walls of the room where the arena was located, to enable easy navigation for the mouse. The arena was illuminated with two 40-W neon bulbs on the ceiling, above the center of the arena. These light conditions were the same for experiments conducted in different phases of the photoperiod cycle.

Tracking was performed with a video camera installed on the ceiling, feeding directly into a personal computer running a Noldus EthoVision video tracking system (Spink, Tegelenbosch, Buma, & Noldus, 2001), using its subtraction mode. Tracking rates and spatial resolutions (the actual distance represented by a single pixel on the screen) are detailed in Table 1. Coordinate files were exported from EthoVision and analyzed by SEE.

Path Analysis

Robust smoothing (i.e., not affected by arbitrary outliers) of the animal path and speed is an important ingredient in the SEE test. We used the Lowess algorithm (Cleveland, 1977), as was described in Kafkafi et al. (2001), with some improvements. The main improvement consisted of adding an algorithm based on repeated running median (RRM; see Tukey, 1977), which was used to smooth the path at the very low velocities without erasing the very short stops (arrests). In a single iteration, the smoothed location for each data point is the median of the locations in a small time window around this point. This process is repeated for several iterations with different window sizes. We used four iterations with window sizes of seven, five, three, and three data points. This choice followed Tukey's (1977) guidelines and a comparison of the outcome with the actual videotaped behavior in several sampled sequences to ensure that short arrests were not smoothed out. The results of the RRM smoothing were used only to isolate arrest intervals, which were defined as instances where the RRM smoothed location did not change for at least 0.2 s, and in these points the speed was defined as zero. Such short stops constitute an important part of the behavioral repertoire of rodents (Golani et al., 1993),

Table 1
Differences in Experimental Groups and Testing Conditions in the Six Experiments

	Across-labs experiments			Within-lab variations		
Parameter	NIDA	MPRC	TAU	MPRC/L	NIDA/LS	NIDA/C
Strains (*n*)	B6 (8) D2 (8)	B6 (10) D2 (10)	B6 (9) D2 (9)	B6 (8) D2 (9)	B6 (5) D2 (6)	B6 (8) CXBK (8)
Laboratory	NIDA	MPRC	TAU	MPRC	NIDA	NIDA
Arena's diameter	250 cm	210 cm	250 cm	210 cm	250 cm	250 cm
Tracking rate	30/s	30/s	25/s	30/s	30/s	30/s
Spatial resolution	1.3 cm	1.0 cm	1.0 cm	1.0 cm	1.3 cm	1.3 cm
Cycle of testing	Dark	Dark	Dark	Light	Light	Light
Treatment					Saline	

Note. The three experiments in the left section were used to test across lab replicability. The three experiments in the right section were used to test the replicability across experiments with different conditions within the same lab, through comparisons with the left-section experiment of the same laboratory. NIDA = National Institute on Drug Abuse; MPRC = Maryland Psychiatric Research Center; TAU = Tel Aviv University; MPRC/L = MPRC/light; NIDA/LS = NIDA/light–saline; NIDA/C = NIDA/CXBK; B6 = C57BL/6; D2 = DBA/2; CXBK = CXBK/ByJ.

especially of the small and fast-moving mouse (Drai et al., 2001), and most other smoothing algorithms, although useful at higher speeds, would tend to smooth them out. For nonzero speeds, the smoothed locations and speed were estimated by the Lowess from the raw data as before (Kafkafi et al., 2001). The time window used for the Lowess was 0.4 s, and the polynomial degree was 2. The above combined procedure was implemented in SEE Path Smoother, a stand-alone program available from us with or independently of the whole SEE package. For a general review of smoothing methods for tracking and discussion of their importance, see Hen, Sakov, Kafkafi, Golani, and Benjamini (2003).

Segmentation of the smoothed path into lingering episodes and progression segments was done with the expectation maximization (EM) algorithm, as in previous studies (Drai et al., 2000; Kafkafi et al., 2001), except for one important difference: The segmentation was always done into two components—lingering and progression—and we did not use the further division of progression into slow and fast movement (e.g., "second and third gears"). The reason is that this subdivision is often not clear in mice, whereas the division into lingering and progression is very general. Most mice displayed a clearly bimodal distribution of segment maximal speeds, with the threshold between the lingering and progression typically at values of 10 to 20 cm/s. As with the smoothing algorithms, the segmentation using the EM algorithm is currently implemented in a stand-alone program, which is available from us with or independently of the whole SEE package.

Visualization, analysis, and calculation of behavioral measures were done with SEE (Drai & Golani, 2001) and with the assistance of two extension programs, the SEE Experiment Explorer and SEE Endpoint Manager (Kafkafi, in press). The first was designed to assist with SEE querying any desired subsection of a database that includes many experiments, whereas the second standardizes SEE calculation of endpoints and the development of new endpoints. These programs are not necessary for the analysis, but they make it much more efficient and user friendly. Both are also available from us. Seventeen behavioral measures (i.e., endpoints) were used in this study and are listed in Tables 2 and 3. In the approach suggested in this study, the algorithms of these endpoints are not merely methods but constitute an important result of the study. The algorithm for generating each endpoint is therefore reported in the *Comparison With a Photobeam Test* section, with its rationale and its success with reliably differentiating the locomotor behavior of B6 and D2.

Statistical Methods

Standard transformations (log, square root, logit) were applied to the results in some of the endpoints (see Table 2) to correct toward approximately normal distributions. Across labs, we analyzed results by comparing Experiments NIDA, MPRC, and TAU using a Genotype × Laboratory two-way analysis of variance (ANOVA) for each endpoint. We estimated the effect of each factor (genotype, laboratory, their interaction, and the residuals, or individual animal) by the proportion of the variation attributed to that factor out of the total variation—that is, SS_{factor}/SS_{total}, and assessed their statistical significance using F tests. We also supported the above by calculating omega-squared estimates for the effects of the factors. In spite of being generally less biased than the proportions of variation, they run into problems when the estimated effects are close to zero, as is the case here. Because elsewhere the values of omega squared were not smaller than the proportion of variation by more than 3%, we report only the latter. Note that the proportion of genotypic variance is a relatively conservative estimate of the broad sense heritability, as some of the interaction variance may also be genetic and part of the individual variance may be attributed to measurement error.

The testing of many endpoints in this study raises the problem of multiple comparisons. We approached the problem using the false discovery rate (FDR), as suggested by Benjamini and Hochberg (1995) and Benjamini and Liu (1999). This approach calls for controlling the expected proportion of false discoveries among the discoveries (the number of erroneously rejected null hypotheses among the rejected ones). When all differences are not real, this protects the experimenter against making even one false discovery, but otherwise FDR controlling procedures are more powerful than traditional multiple comparison procedures. See Benjamini et al. (2001) for a detailed discussion of the approach with behavioral phenotyping and explanation of the procedures used to assess the significances obtained from the two-way ANOVA. The FDR was controlled separately for genotype, lab, and interaction p values, each at a level of .05.

We tested the replicability across experiments with different conditions within the same lab by comparing Experiment NIDA with Experiment NIDA/LS and comparing Experiment MPRC with Experiment MPRC/L (see Table 1), using the same procedures as for the three lab comparisons.

Results

Table 2 presents the results in each endpoint (group means and standard deviations) of the two genotypes in the three labs, NIDA, MPRC and TAU, and in the two additional experiments, MPRC/L and NIDA/LS (see Table 1 for locations and conditions). Table 3 presents the genotype differences in each endpoint using Genotype × Lab two-way ANOVA as tested across NIDA, MPRC, and TAU and corrected for multiple comparison by FDR. Significant strain differences were found in 15 out of 17 endpoints. Out of these 15 endpoints, 7 had significant lab effects. The lab effects were much smaller than the genotype effects, except for home base relative occupancy (see next section for the description of endpoints). None of the endpoints, however, had a significant Genotype × Lab interaction. This means that lab effects were all additive. That is, the genotype differences did not differ across labs even when the actual group means did. Regarding effect sizes, in 12 out of the 17 endpoints, the genotypic variance was larger than the laboratory and interaction variances combined. In 7 out of these endpoints, the genotypic variance alone accounted for more than 50% of the total variance.

Experiment MPRC/L replicated the experiment in the MPRC, except that animals were tested during their light phase instead of their dark phase. The two experiments were compared with a Genotype × Experiment two-way ANOVA. All the genotype differences that were significant in the across-labs comparison were also significant in this comparison except for one: the diversity. The only endpoint that was significantly different between the two experiments was the lingering spatial spread. None of the interaction effects was significant.

Experiment NIDA/LS replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. Out of the 15 genotype differences that were significant in the across-labs comparison, only 4 were not significant in this comparison: the maximal segment speed, the diversity, the number of stops per excursion, and the home base relative occupancy. The only endpoint that was significantly different in Experiment NIDA/LS was the latency to half maximum speed. As in the previous comparisons, none of the interaction terms was significant.

Note that the two within-lab comparisons above do not appropriately test the effect of the differences in their protocol, the testing phase and the saline injection, as these differences were not controlled factors within the same experiment but differences of conditions between two experiments conducted at different occasions. Moreover, in principle, the effect of repeating the experi-

Table 2
Strain Means in the Three Laboratories (NIDA, MPRC and TAU) and for Experiments NIDA/LS and MPRC/L

		NIDA		MPRC		TAU		NIDA/LS		MPRC/L	
Endpoint	Units/transform	B6	D2	B6	D2	B6	D2	B6	D2	B6	D2
1. Distance traveled	m										
M		256	144	227	130	264	109	234	79	247	181
SD		42	54	13	48	40	40	33	54	24	36
2. Center time	Proportion/logit										
M		−0.89	−3.20	−0.99	−3.40	−0.79	−3.81	−0.54	−3.60	−0.60	−2.20
SD		0.25	1.06	0.16	1.87	0.47	1.08	0.49	2.10	0.38	0.32
3. Proportion of lingering	Proportion										
M		0.49	0.80	0.50	0.80	0.30	0.70	0.53	0.87	0.47	0.67
SD		0.10	0.10	0.03	0.10	0.03	0.10	0.07	0.09	0.06	0.06
4. Number of progression segments	Number										
M		554	286	593	317	525	223	479	192	575	448
SD		67	87	73	118	50	102	124	144	98	80
5. Lingering mean speed	cm/s										
M		3.60	2.30	2.90	2.00	2.00	1.50	3.49	1.48	3.10	1.90
SD		0.40	0.70	0.40	0.40	0.20	0.50	0.32	0.57	0.43	0.18
6. Lingering spatial spread	cm/square root										
M		1.34	1.51	1.28	1.60	1.00	1.36	1.46	1.53	1.19	1.24
SD		0.18	0.29	0.07	0.21	0.06	0.35	0.21	0.15	0.13	0.06
7. Length of progression segments	cm										
M		28.9	29.9	23.4	24.6	28.0	31.7	31.8	27.4	26.8	26.2
SD		7.7	9.9	4.3	5.0	5.1	12.8	8.8	7.6	4.4	5.4
8. Diversity	Number										
M		132.0	120.0	116.0	111.0	140.0	120.0	123.6	91.4	115.9	114.8
SD		6.9	29.3	2.5	14.9	5.9	36.9	15.8	54.7	3.3	7.1
9. Maximal segment speed	cm/s										
M		31.0	38.9	27.6	32.9	29.0	34.8	30.2	29.5	29.6	33.7
SD		3.6	7.7	2.9	4.4	4.1	8.6	3.6	6.4	3.8	4.1
10. Segment acceleration	cm/s^2/log										
M		3.2	3.6	3.2	3.5	3.0	3.2	3.1	3.3	3.2	3.4
SD		0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.4	0.1	0.1
11. Rate of turn	deg/s										
M		25.6	20.8	31.6	24.6	28.2	19.0	25.8	19.1	30.6	24.2
SD		1.4	3.1	2.4	3.1	2.7	1.9	0.3	1.9	1.7	2.1
12. Radius of turn	Ratio/log										
M		−0.62	−0.12	−0.74	−0.26	−0.80	−0.16	−0.68	−0.33	−0.71	−0.30
SD		0.07	0.11	0.11	0.17	0.14	0.12	0.09	0.24	0.10	0.08
13. Home base relative occupancy	Ratio/log										
M		2.0	2.3	1.1	1.6	1.6	2.0	2.0	2.5	1.2	1.5
SD		0.3	0.5	0.2	0.3	0.4	0.5	0.5	0.3	0.2	0.3
14. Number of excursions	Number/square root										
M		6.0	4.5	5.5	4.4	5.3	3.6	5.9	3.5	5.3	4.9
SD		0.5	0.8	0.5	0.9	0.4	0.9	0.7	1.4	0.2	0.5
15. Number of stops per excursion	Number										
M		11.8	8.9	13.9	8.2	12.1	11.4	10.2	10.8	14.5	13.7
SD		2.1	3.4	4.3	3.9	4.1	3.1	2.5	2.9	2.1	2.8
16. Activity decrease	m										
M		−13.0	−26.0	−40.0	−20.0	−17.0	−21.0	9.6	21.1	−21.2	−32.8
SD		13.0	20.0	16.0	16.0	22.0	23.0	6.5	13.1	17.6	19.0
17. Latency to half maximum speed	s/log										
M		2.9	4.5	3.2	4.2	3.6	4.9	4.3	5.7	3.6	4.3
SD		0.4	0.4	0.8	1.1	0.5	0.5	0.5	1.2	0.9	0.5

Note. Experiment MPRC/L replicated the experiment in the MPRC, except that animals were tested during their light phase. Experiment NIDA/LS replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. For additional differences in experiment conditions, see Table 1. In endpoints in which transformations were used for the statistical analysis, the means and the standard deviations are for the transformed variables. NIDA = National Institute on Drug Abuse; MPRC = Maryland Psychiatric Research Center; TAU = Tel Aviv University; NIDA/LS = NIDA/light–saline; MPRC/L = MPRC/light; deg = degrees.

ment might have countered the effect of the different conditions in the second experiment, to create a false impression of replicability. Genotype differences that were significant in both within-lab comparisons, however, were also significant in the across-labs comparison. Together with the small size and lack of significance of the interaction terms, this strongly suggests that the genotype differences were replicated in the different experiments despite the differences in laboratory and conditions.

Table 3
Effect Sizes as Estimated by Percentage of Variance (% V) Attributable to Genotype, Laboratory, and Interaction in Each Endpoint and Their Statistical Significances

	Genotype			Laboratory			Interaction		
Endpoint	$F(1, 48)$	p	% V	$F(2, 48)$	p	% V	$F(2, 48)$	p	% V
1. Distance traveled	118.4	.0001*	68.3	1.1	.335	1.3	2.4	.1055	2.7
2. Center time	86.5	.0001*	63.6	0.2	.827	0.3	0.6	.575	0.8
3. Proportion of lingering	337.2	.0001*	76.6	26.8	.0001*	12.2	0.7	.494	0.3
4. Number of progression segments	141.2	.0001*	71.6	3.8	.0294	3.9	0.2	.8086	0.2
5. Lingering mean speed	61.3	.0001*	33.3	33.3	.0001*	36.3	3.9	.0227	4.3
6. Lingering spatial spread	34.1	.0001*	31.7	11.2	.0001*	20.9	1.4	.25	2.7
7. Length of progression segments	0.6	.444	1.1	3.3	.0471	11.7	0.2	.793	0.8
8. Diversity	5.2	.0274	8.5	3.3	.0448	10.8	0.7	.5246	2.1
9. Maximal segment speed	15.8	.0002*	22.5	2.9	.0656	8.2	0.3	.715	1.0
10. Segment acceleration	114.7	.0001*	51.2	29.0	.0001*	25.9	1.7	.187	1.5
11. Rate of turn	106.6	.0001*	52.1	21.6	.0001*	21.1	3.4	.0403	3.4
12. Radius of turn	247.1	.0001*	79.8	5.0	.0105	3.2	2.3	.1101	1.5
13. Home base relative occupancy	12.2	.001*	11.9	21.0	.0001*	40.8	0.3	.7095	0.7
14. Number of excursions	54.0	.0001*	47.0	5.7	.0062*	9.9	0.7	.483	1.3
15. Number of stops per excursion	9.9	.003*	15.6	0.7	.4908	2.3	2.1	.138	6.5
16. Activity decrease	0.2	.67	0.3	2.0	.1486	6.6	3.8	.0295	12.7
17. Latency to half maximum speed	46.7	.0001*	44.7	4.4	.0183	8.3	0.5	.5813	1.0

Note. Effects denoted by asterisks were found to be significant when the false discovery rate was controlled at .05. Note that the percentage of genotype variance is a conservative estimate of broad-sense heritability.

All the endpoints except for the last two, activity decrease and the latency for half maximal speed, can be measured in time bins of 5 min. Endpoints that discriminated the two strains over the whole session were mostly able to discriminate them in each of the 5-minute bins, as is demonstrated in Figures 1, 2, and 3 for distance traveled, center time, and radius of turn. Note that the development of the radius of turn seems to differentiate the CXBK mice in Experiment NIDA/C from the other two strains. The CXBK started the session like the B6 and ended like the D2. In general, note that group variances were usually small even when measured in bins of 5 min, although group sizes were not large (5 to 10).

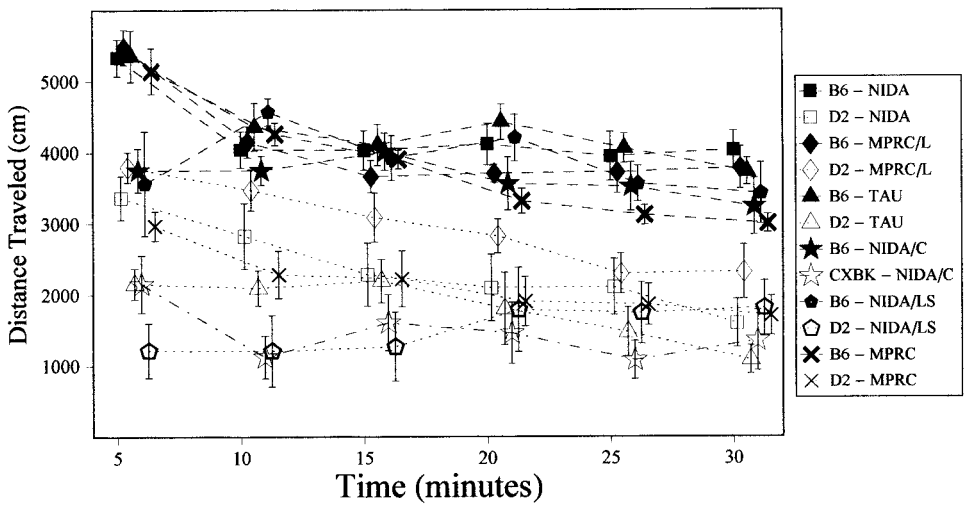

Figure 1. The distance traveled in the six experiments (group means plus or minus standard error) in time bins of 5 min. Dark symbols with dashed lines represent C57BL/6 (B6) groups. Open symbols with dotted lines represent DBA/2 (D2) groups. An open star with dash–dot line represents a single CXBK/ByJ (CXBK) group. Experiment Maryland Psychiatric Research Center/Light (MPRC/L) replicated the experiment in the MPRC, except that animals were tested during their light phase instead of their dark phase. Experiment National Institute on Drug Abuse/Light–Saline (NIDA/LS) replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. Experiment NIDA/C compared B6 with CXBK mice in NIDA with animals tested during the light phase. For additional differences in experiment conditions, see Table 1.

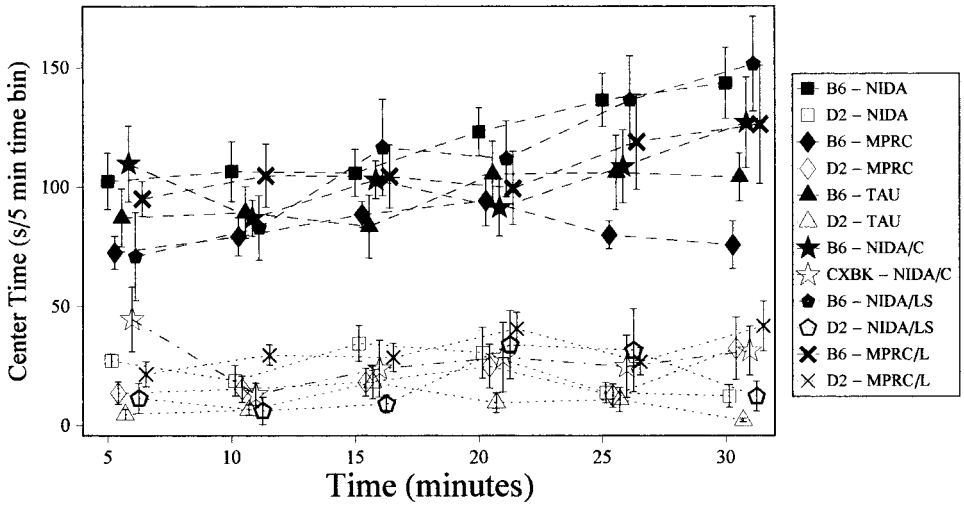

Figure 2. Center time in the six experiments (group means plus or minus standard error) in time bins of 5 min. Dark symbols with dashed lines represent C57BL/6 (B6) groups. Open symbols with dotted lines represent DBA/2 (D2) groups. An open star with dash–dot line represents a single CXBK/ByJ (CXBK) group. Experiment Maryland Psychiatric Research Center/Light (MPRC/L) replicated the experiment in the MPRC, except that animals were tested during their light phase instead of their dark phase. Experiment National Institute on Drug Abuse/Light–Saline (NIDA/LS) replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. Experiment NIDA/C compared B6 with CXBK mice in NIDA with animals tested during the light phase. For additional differences in experiment conditions, see Table 1. TAU = Tel Aviv University.

Endpoint List

Endpoint algorithms constitute an important result of this study, as replication of genotype differences across laboratories and conditions with a certain endpoint suggests that the algorithm of this endpoint captures a genotype-specific behavior pattern. In what follows, we describe each endpoint separately, with the rationale of using it, and present its success in reliably differentiating the

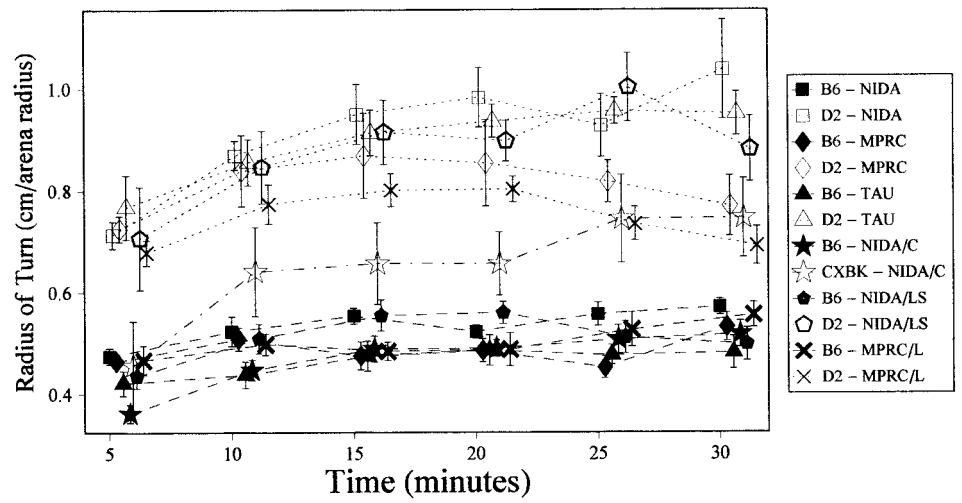

Figure 3. Radius of turn during progression in the six experiments (group means plus or minus standard error) in time bins of 5 min. Dark symbols with dashed lines represent C57BL/6 (B6) groups. Open symbols with dotted lines represent DBA/2 (D2) groups. An open star with dash–dot line represents a single CXBK/ByJ (CXBK) group. Experiment Maryland Psychiatric Research Center/Light (MPRC/L) replicated the experiment in the MPRC, except that animals were tested during their light phase instead of their dark phase. Experiment National Institute on Drug Abuse/Light–Saline (NIDA/LS) replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. Experiment NIDA/C compared B6 with CXBK mice in NIDA with animals tested during the light phase. For additional differences in experiment conditions, see Table 1. TAU = Tel Aviv University.

two genotypes. The endpoints measure the behavior of an individual mouse over a session. Many of the endpoints quantify properties of segments (either progression or lingering episodes) performed by the mouse throughout the session, and in these cases the median over all segments in the session was taken, unless mentioned otherwise.

1. Distance traveled. This is the total length of the smoothed path. This measure might seem trivial, as it is the primary endpoint used in most locomotor tests. In most of these tests, however, it is highly sensitive to the spatial and temporal resolution of the tracking because higher resolution can detect smaller meanderings of the path and thus increase its length considerably (Paulus & Geyer, 1993). Consequently, travel distances measured with different tracking parameters and arena sizes are usually not considered to be comparable. In addition, this measure is sensitive to recognition errors of the tracking system. For example, if the system erroneously identified the animal as being in the other side of the arena for only a single frame, the distance from the true location and back is added to the distance traveled, and such an error might happen many times during a session. The robust smoothing algorithms used in our method, however, take care of both these problems. In addition, the large arena may cause the B6 to increase their mileage and/or cause the D2 to decrease their mileage. This may explain why B6 mice traveled very significantly larger distances, about 250 m versus 130 m in our setup. Furthermore, the strain means themselves, not only the difference between them, were very consistent both across and within labs, despite the differences in arena size, protocol conditions, and tracking parameters.

2. Center time. This endpoint measures the total time the animal spent at a distance of more than 15 cm from the wall of the arena. B6 mice spent significantly more time in the center than did D2 mice. This difference as well as the strain means themselves were very replicable, both across labs and across experiments with different conditions within labs, as shown by less than 1% of the total variation attributable to any lab, experiment, or interaction effect, none of which we found statistically significant (see Table 3 and Figure 2). Such a difference was not reported for these two genotypes with smaller arenas, although most current photobeam systems measure center time regularly (see Figure 4, middle row). Because center time, in contrast to distance traveled, is not very sensitive to tracking resolution or tracking artifacts, it seems that the large arena used in our protocol accentuated the differences in wall hugging.

3. Proportion of lingering time. The proportion of lingering time equals the total duration of lingering episodes (i.e., stops, as computed by the segmentation EM algorithm) in the session as a proportion of the session duration. D2 mice had much longer lingering time, about 30% to 40% more than B6, the difference being very significant. The lab effect was significant, but within-lab differences were not.

4. Number of progression segments. The number of progression (movement) segments in the 30-min session, as computed by the segmentation process, is, by definition, also the number of stops. B6 mice had significantly higher numbers of progression segments than D2 mice. Strain differences as well as strain means themselves were replicable across experiments with different conditions within labs (see also Figure 4, bottom row).

5. Lingering mean speed. The lingering mean speed equals the cumulative distance traveled in the lingering (stopping) mode, divided by the cumulative duration of lingering. It thus provides a rough measure of local mobility within stops, which probably consists mainly of scanning movements, few sideways and forward steps, rearing, and stretch-attend behavior. B6 mice had higher lingering mean speed than D2 across and within laboratories. The large and very significant lab effect in this endpoint might reflect the sensitivity of small movement measurement to tracking conditions and parameters. This is also suggested by the lack of significant effect in the comparisons within laboratories. A similar pattern was also found with the proportion of lingering time (see the *Proportion of lingering time* section). As with all other comparisons, however, the Genotype $\times$ Lab interaction was not significant.

6. Spatial spread of lingering episodes. The spatial spread of lingering episodes is the longest distance between any two points during a lingering episode (a stop). Because lingering episodes are frequently local and circumscribed, their spatial spread is a more appropriate measure than their length (Drai et al., 2000). D2 had a significantly larger spatial spread. The significant lab effect may also be due to the lower reliability of tracking small lingering movements, but the interaction effect was not significant. Within the MPRC, the spatial spread was significantly smaller in Experiment MPRC/L, in which the animals were tested during the light cycle of their photoperiod.

7. Length of progression segments. Regarding the path length of progression (movement) segments, no differences were found between the two genotypes. All group means were between 23 and 33 cm. This similarity seems to be a property of the B6 and D2 strains, as some of the additional strains we currently test do show significantly shorter or longer progression segments.

8. Diversity. Diversity is the average distance between any two stops, weighted by the distribution of the duration of these stops (Tchernichovski et al., 1996). This measure captures the spatial and temporal scatter of stops. It is higher as a greater area is covered, as it is covered more homogenously, and as the duration of stopping is distributed more homogeneously over the arena. B6 had significantly higher diversity, but the proportion of variance attributed to the individual animal in this endpoint was very large, about 79% (see Table 3), and the strain difference was not significant in either within-lab comparison.

9. Maximal speed of progression segments. The maximal speed attained during a segment was found to distinguish progression from lingering behavior (Drai et al., 2000; Kafkafi et al., 2001) and is therefore a reasonable choice for genotype discrimination. The median of segment maxima over all progression segments in the session was taken as the result of this animal. Despite the much higher distance traveled by B6 mice, D2 mice were significantly faster across labs with no significant lab effect. This difference was significant also in the comparison within the MPRC but not in the comparison within NIDA.

10. Segment acceleration. Segment acceleration provides, for each progression segment, its maximal speed divided by its duration. It is thus a rough estimation of the acceleration in this segment (Kafkafi et al., in press). D2 had significantly higher segment acceleration than B6 across labs. The laboratory effect was also significant. In both comparisons within labs, genotype

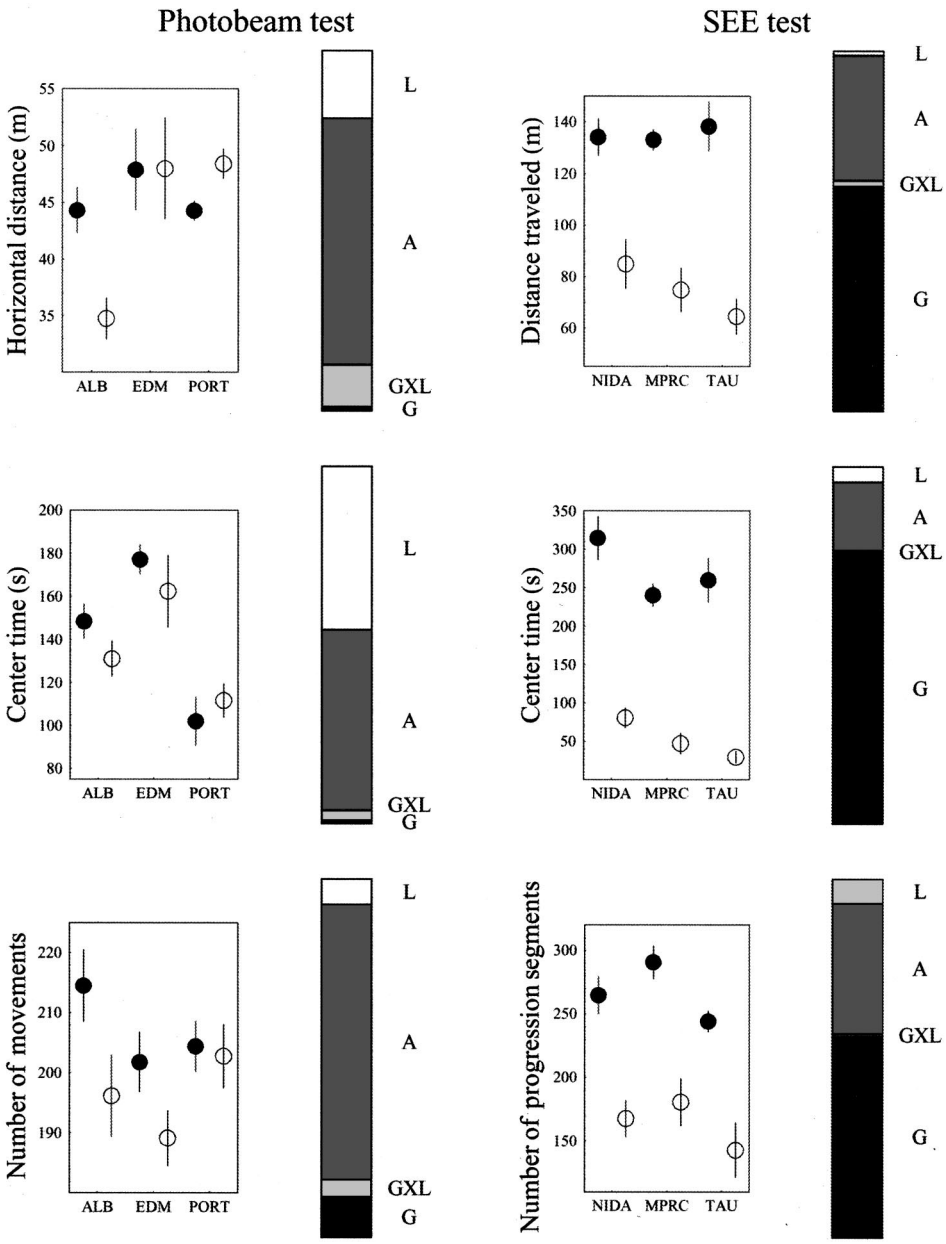

Figure 4. Comparison between a photobeam test (left) and the Strategy for Exploring Exploration (SEE) test (right) of the differences detected between C57BL/6 (B6; closed circles) and DBA/2 (D2; open circles) across three laboratories in three corresponding analogous endpoints: horizontal distance versus distance traveled (top), center time versus center time (middle), and number of movements versus number of progression segments (bottom). Error bars represent standard errors. Photobeam results are from a study described in Crabbe et al., 1999b. The laboratories in Crabbe et al. were Albany (ALB), Edmonton (EDM), and Portland (PORT). Laboratories in the SEE test were National Institute on Drug Abuse/IRP Laboratory (NIDA), Maryland Psychiatric Research Center (MPRC), and Tel Aviv University (TAU). Columns at the right represent the proportion of variance attributed to each factor. L = laboratories; A = individual animals; G = genotypes; G × L = Genotype × Laboratory interaction. Note that the interaction parts G × L in the center time and number of progression segments in the SEE test (middle and bottom right) are too small to be noticed in their columns. Data for both tests are based on 15-min session duration and $n = 8$ males, and no transformations are used.

effects were significant but experiment and interaction effects were not.

11. Rate of turn during progression. The rate of turn during progression measures the amount of turning (change of progression direction) in time during progression segments. For each data point in the progression mode, this endpoint calculates the change of heading relative to the previous data point. Division by the tracking rate gives the rate of direction change in degrees per second for each data point. The session value is the median of the absolute rate (i.e., no distinction between right and left turning) over progression during the whole session. This measure is not computed in the lingering mode, as the distance between consecutive data points in this mode might be very small, frequently much less than the spatial resolution of the tracking system, and thus the measured change in direction is likely to be meaningless. B6 rate of turn was very significantly higher than that of D2 across laboratories. Laboratory effect was also significant, but in both within-lab comparisons, genotype differences were significant, whereas experiment and interaction effects were not.

12. Radius of turn during progression. This measure is computed by dividing, for each data point, the absolute turning rate (in degrees/s) by the speed (in cm/s) to get the curvature (in degrees/cm). The result is further multiplied by $180/\pi$ to get the radius of turning in centimeters. This is a measure of turning similar to the previous, but in relation to space and not to time. Note that it is possible to turn with the same radius but with different turning rates (by changing the speed) or with different turning radii but with the same turning rate. As with the turning rate, we used the median over all data points during the progression mode only. Furthermore, because one of our labs had an arena with a slightly smaller radius and the animals progressed mainly along the wall, we calibrated the median radius of turn for each session by dividing it by the arena radius. The radius of turn (Figure 3) of the B6 was very significantly smaller in both across- and within-lab comparisons, about half of the arena radius, whereas the radius of turn of the D2 was almost as large as the arena radius. Note that the proportion of variance attributed to the genotype, which is a conservative estimation of broad sense heritability, was almost 80% in this endpoint, whereas merely 3.2% and 1.5% of the variance were attributed to lab and interaction effects, respectively.

13. Home base relative occupancy. Home base relative occupancy measures how much the animal stays in its most preferred location. This endpoint is computed by multiplying the total lingering duration by the number of lingering episodes for each place in the arena. This calculation is performed in bins of 10° along the perimeter of the arena. The home base was defined as the bin having the maximum value of this multiplication. It was usually, but not always, the place where the mouse was first introduced. The occupancy of the home base is given as the ratio between this maximum and the mean over all 10° bins. Note that occupancy in this algorithm uses both the time of staying and the number of stops, as found by Eilam and Golani (1989). D2 mice had significantly higher home base occupancy across laboratories. This endpoint was, however, the only one in which the lab effect was much larger than the genotype effect. The genotype differences were also significant in the within-MPRC comparison but not in the within-NIDA comparison.

14. Number of excursions. A round trip starting and ending at the home base is an excursion, which is a natural unit of rodent exploration (Eilam & Golani, 1989; Golani et al., 1993; Tchernichovski et al., 1998; Tchernichovski & Golani, 1995; also recently used in Whishaw et al., 2001; and Wallace et al., 2002). B6 mice had significantly more excursions, about 30 versus 20 by the D2 mice, both across and within laboratories. The lab effect was also significant, but the size of its effect was about a fifth of the strain effect.

15. Number of stops per excursion. The mean number of stops in each excursion was found to be useful with Long–Evans rats (Golani et al., 1993). It was also used to distinguish between rats showing high and low response to novelty under dexamphetamine (Cools et al., 1997; Gingras & Cools, 1997). Session median was typically between 8 and 15, which is perhaps slightly larger than that of rats, and B6 had significantly more stops per excursion across laboratories. This difference was also significant in the comparison within the MPRC but not in the comparison within NIDA.

16. Activity decrease. Activity decrease measures the difference in activity between the first and second 15 min of the session. No significant differences were found between B6 and D2 in this regard. It is interesting that this measure yielded the most significant difference between B6 and BALB/c strain (a single-lab study, Benjamini et al., 2001).

17. Latency to half of maximal speed. This endpoint quantifies how fast the activity of the animal builds up during the beginning of the session. It measures the time, from the start of the session, it took the animal to exceed for the first time a speed equaling half of the maximal speed it attained during the whole session. B6 mice were significantly faster to take off, both across and within labs, typically around 10 s, whereas D2 typically took around 100 s. In Experiment NIDA/LS this endpoint was significantly higher, as compared with the other experiment in this lab. This was probably a result of the saline injection in NIDA/LS, as no such change was detected in the comparison within the MPRC.

Comparison With a Photobeam Test

The multilaboratory approach first taken by Crabbe et al. (1999a) prompted our efforts to explore improved tests as a mean for achieving higher replicability. To this end, we have furthered their discussion by directly comparing the replicability of our results with the replicability of the locomotor behavior results measured by AccuScan Digiscan photobeam systems in complete darkness, as published in their Website at http://www.albany.edu/psy/obssr3/. This is especially needed because the SEE test is not yet as standardized and commercialized as the photobeam tests, requiring both more space to be conducted and motivation for replacing the familiar with the less familiar. It is thus reasonable to ask whether the gain in performance is large enough to justify such an investment. Moreover, the SEE test cannot be said to be more replicable than the photobeam system merely because this study did not find significant Genotype $\times$ Lab interactions, whereas Crabbe et al. (1999a) did. This difference could be attributed to a higher power of the second study to detect interactions, as it included more genotypes and larger (when collapsed over sex) group size.

To equalize strains, sexes, session duration, and group size, we used only the photobeam test B6 and D2 males in all three laboratories, only the first 15 min of each SEE test session, and

only 8 randomly chosen animals from each SEE test group. We compared the genotype differences found in three corresponding analogous endpoints: horizontal distance, center time, and number of movements in the photobeam system with the respective distance traveled, center time, and number of progression segments in the SEE test. No transformations were used in this comparison with any of the endpoints.

Figure 4 displays the results of this comparison, with the columns representing the proportion of variance attributed to genotype, laboratory, the Genotype × Laboratory interaction, and individual animals. The variances attributed to the genotype are seen to be large for all three measures in the SEE test and are highly statistically significant ($p < .0001$). The variance attributed to the interaction in the SEE test is too small to estimate graphically in Figure 4 (all ps $> .29$). It was 1.2% in the distance traveled (vs. 11.5% in the photobeam test's horizontal distance, where $p < .05$), 0.54% in the center time (vs. 2.57 % in the photobeam test's center time), and 0.59% in the number of progression segments (vs. 4.6% in the photobeam test's number of movements). The lab effects were also smaller in the SEE test than in the photobeam test and generally less significant (in the SEE test, $p \leq .05$ for center time and number of progression segments, $p > .49$ for distance traveled; in the photobeam test, $p < .01$ for horizontal distance, $p \leq .0001$ for center time, and $p > .16$ for number of movements). The locomotor behavior test with SEE analysis thus increased genotypic effect and decreased laboratory, interaction, and individual effects several times over. Note that, in contrast with the photobeam test, this advantage was achieved without any special effort to equalize housing and testing conditions (see Table 1).

The third comparison of the photobeam system's number of movements with SEE's number of progression segments (Figure 4, bottom two graphs) also illustrates some of the problem with using significance in the two-way ANOVA for assessing across-labs replicability. The genotype difference in the photobeam system is actually significant, whereas the laboratory and interaction effects are not. With SEE, the genotype effect is much more significant, but the lab effect is also marginally significant. This is mainly due to the higher discrimination power of SEE, as is evident from smaller group variances in the graph and smaller proportion of individual variance in the column. Because F ratios of all factors in the standard ANOVA are computed relative to this pooled within-group variance, SEE did better with detecting the genotype difference but also with detecting the lab difference. Assessing replicability by lab and interaction significance without considering also the genotype significance thus penalizes methods with higher discrimination power, and in this case, might have led one to believe that SEE analysis was less replicable across laboratories, although the graphs show clearly that the opposite is true.

It is important to note that by comparing the performances in endpoints such as horizontal distance in the photobeam system and distance traveled in the SEE test, we do not imply that these endpoints necessarily measured the same behavior in the two tests, in light of the considerable differences in test conditions. In fact, the photobeam results included many additional endpoints, such as vertical activity and number of movement, number of clockwise and counterclockwise revolutions, and stereotypy count. None of these, however, detected much larger genotype differences than the three endpoints shown in Figure 4. The only differences that were highly replicable were the advantage of the B6 in several endpoints, such as the horizontal distance and number of movements, during the first 5 min of the session only. The same pattern can be seen with the distance traveled measured in a single laboratory experiment by Jones et al. (1993). This agrees with our results in the measure of latency to half maximum speed, suggesting that B6 are faster starters than D2. It appears that in the photobeam systems, this initial advantage of the B6 disappears after the first 5 min, when the two strains both covered the entire small arena. In the large arena used in this study, however, genotype differences were generally maintained, as is illustrated in Figures 1, 2 and 3, even after 30 min.

Discussion

The B6 and D2 are two of the most commonly used inbred strains of laboratory mice. Consequently, many behavioral differences between them have been reported in the literature. The results of most reports, however, have not been methodically corroborated in more than one laboratory. The findings of Crabbe et al. (1999a) suggest that such differences could well prove to be idiosyncratic to the specific laboratory. The purpose of the experiments described in this report was to explore the heritability and replicability of endpoints derived from ethologically based studies of exploratory behavior (Benjamini et al., 2001; Drai et al., 2000, 2001; Eilam & Golani, 1989; Eilam et al., 1989; Golani et al., 1993; Kafkafi et al., 2001, in press; Tchernichovski et al., 1998).

Replicable differences between the two genotypes were found in several different aspects of locomotor behavior. B6 mice traveled longer distances, performed more stops (and, by definition, more progression segments), and spent less time in lingering. The B6 mice increased their distance traveled by increasing the number of progression segments, not by increasing the typical length of progression segments. D2 mice, however, had higher maximal speeds and accelerations. This combination of differences is counterintuitive to a general activity view of locomotor behavior. According to such a view, optional measures of activity, such as the distance traveled or the number of beam breaks in different photobeam arrangements, are most likely to be correlated, and thus the exact definition of an activity measure is not that crucial. Activity, as measured by a variety of hardware and algorithms, is the endpoint reported in most locomotor behavior studies and is used to assess the effect of many drugs and treatments. Our results suggest, however, that it artificially combines several different aspects of behavior that are not necessarily correlated in a trivial way.

B6 mice spent a much longer time away from the wall than did D2 mice. They typically took only several seconds from the start of the session to reach half of the maximal speed attained in the session, whereas D2 typically took a minute or so. B6 mice's progression was less straight, in relation to both space (turning radius) and time (turning rate). Most differences found over the whole 30-min session were also prominent in most of the 5-min time bins of the session.

A conservative estimate of the broad-sense heritability was higher than 50% with 7 endpoints out of the 17 tested in this study, despite the fact that the tests were conducted over three laboratories with slightly different conditions. Two complementary approaches for mapping genes affecting behavioral traits are quantitative trait loci (QTL, e.g., Belknap at al., 2001; Flint et al., 1995;

Turri, Talbot, Radcliffe, Wehner, & Flint, 1999) and mutagenesis (e.g., Nolan et al., 2000; Takahashi, Pinto, & Vitaterna, 1994). Mutants and recombinant inbred strains (the BXD strains) derived from B6 and D2 play important role in both QTL and mutagenesis studies conducted in mice. With both approaches, the discrimination power and replicability of phenotyping crucially affect the quality of the results. Both approaches may thus gain from the use of the SEE test.

It is interesting that mice of both strains were not more active when tested during the dark phase of the photoperiodic cycle than during the light phase. This, however, may be a result of the light in the testing arena. The contrast between the dark housing room and the lighted arena might have had some inhibiting effect that balanced the higher activity expected during the dark phase. Note that it is not possible to draw conclusions about the effect of the photoperiod phase or saline injection, as they were not manipulated within a single experiment in a single laboratory. The overall results strongly suggest, however, that a considerable proportion of our endpoints were much more sensitive to genotype differences than to conditions such as photoperiod phase and saline injections.

Because all experiments in this study included only two strains each, it was not possible to measure correlations between endpoints across strains. We are currently studying this issue in a multilab experiment involving 10 strains. Although we expect certain endpoints to be usually correlated, the particular behavioral phenotype that characterizes certain genotypes may well be the absence of such a correlation. The CXBK results from Experiment NIDA/C supports this possibility. For example, it can be argued that the radius of turn during progression (Figure 3) is probably inversely correlated with the center time (Figure 2), as animals walking mostly along the wall are likely to have a progression radius similar to the arena radius, and, indeed, the D2 mice, which spent most of their time near the wall, also had a progression radius consistently close to the arena radius, whereas the B6 mice, which spent more time in the center, consistently had a progression radius of about half the arena radius. Hypothetically, however, an animal can walk near the wall in tight arcs or travel through the center in straight progression segments. Note that the center time of the CXBK group in Experiment NIDA/C seems not to differ from that of the D2 groups in the other experiments (Figure 2), but their radius of turn was much smaller than that of the D2 during the first half of the session (Figure 3), closer to that of the B6.

Multilab experiments are currently assessed using the standard method of Genotype $\times$ Laboratory two-way ANOVA. In this model, large size and high significance of lab effects, and even more so of the Genotype $\times$ Laboratory interaction effects, are considered as indicating a replicability problem (Crabbe et al., 1999a; Wahlsten, 2001). Such significance, however, should not be considered separately from the size and significance of the genotype effect. Methods with higher discrimination power are likely to increase significance of all effects because of the small within-group variance. Good replicability should thus be indicated by a proportion of variance due to genotype that is several times larger than that due to the laboratory and especially the interaction (e.g., Figure 4). This problem may be solved by the use of a mixed model instead of a fixed model ANOVA, with the laboratories (and therefore also the interaction) regarded as a random factor. We are currently engaged in adapting this strategy to the problem at hand.

In general, the SEE locomotor behavior test was much more sensitive to the differences between B6 and D2 than was the standard locomotor photobeam test, as demonstrated here by a direct comparison with data published from a multilab study by Crabbe et al. (1999a), whereas the sensitivity to the laboratory effect and Genotype $\times$ Laboratory interaction were generally lower. It is difficult to determine how much of the increased performances should be attributed to the different setup, to the different tracking system, to the robust smoothing, or to the more sophisticated analysis. With most endpoints it was probably an effect of more than one of the above factors and their interaction. In addition to the much larger arena in this study (210–250 cm vs. 45 cm), it is also important to note that the animals in the photobeam test were run in complete darkness, as opposed to usual room lights in the SEE setup. In the case of the center time especially, it is probable that both the large arena and the lights were the main factors that uncovered the big difference between the strains. With rats, we noticed that darkness typically lowers the range of speeds used by the animal considerably (Benjamini et al., 2001), thus limiting the its behavioral repertoire and lowering the quality of SEE analysis, which is highly dependent on the speed. Arena size and light level, as most properties of the setup in this study, were deliberately chosen to increase the resolution and the performance of the analysis.

To encourage the use of this new test, we developed a simple stand-alone program that automatically performs the analysis described in this article, including the smoothing, segmentation, and computation of all the endpoints described above. This program, as is the SEE notebook itself, is available from us. It of course remains to be shown that the high performances of the SEE test can be generalized to other common and important genotypes, an issue that we are currently studying. Note also that SEE analysis can be easily upgraded. New endpoints can be readily developed in SEE, and current endpoints can be updated, on the basis of ongoing results (Kafkafi et al., in press). In our experience, the study of additional strains tends to promote this process because each new strain displays more clearly some patterns that highlight additional properties of locomotor behavior. The performance of the SEE test is thus likely to be further improved in the future by the diverse efforts of the research community.

Standardization is the usually advocated remedy for the problem of phenotyping replicability across laboratories. It is estimated, however, to require a substantial and highly coordinated effort by many laboratories (Wahlsten, 2001), and the level of standardization needed for satisfactory replicability is yet to be demonstrated. Moreover, behavioral tests that are sensitive to any small change in protocol, even within the same laboratory, are likely to make research much more difficult. In contrast, the high replicability of the SEE test in this study was achieved despite several differences in tracking parameters and only mild standardization of housing conditions. Within labs, experiments conducted during the light instead of the dark cycle and with saline injections did not blur most genotype differences and did not significantly interact with any of them. This suggests that tests and measures that are specifically designed to capture ethologically relevant behavior patterns are likely to be more resistant to environment manipulations and thus constitute a fruitful approach for behavioral phenotyping.

References

Belknap, J. K., Hitzemann, R., Crabbe, J. C., Phillips, T. J., Buck, K. J., & Williams, R. (2001). QTL analysis and genomewide mutagenesis in mice: Complementary genetic approaches to the dissection of complex traits. *Behavior Genetics, 31*(1), 5–15.

Benjamini, Y., Drai, D., Elmer, G. I., Golani, I., & Kafkafi, N. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research, 125,* 279–284.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57,* 289–300.

Benjamini, Y., & Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference, 82*(1–2), 163–170.

Bolivar, V. J., Caldarone, B. J., Reilly, A. A., & Flaherty, L. (2000). Habituation in the open field: A survey of inbred strains and F1 hybrids. *Behavior Genetics, 30,* 258–293.

Cabib, S. (2002). The contribution of studies in inbred strains of mice to understanding of a hyperactive phenotype. *Behavioural Brain Research, 130,* 103–109.

Cabib, S., & Bonaventura, N. (1997). Parallel strain-dependent susceptibility to environmentally-induced stereotypies and stress-induced behavioral sensitization in mice. *Physiology & Behavior, 61,* 499–506.

Cleveland, W. S. (1977). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistics Association, 74,* 829–836.

Cools, A. R., Ellenbroek, B. A., Gingras, M. A., Engbersen, A., & Heeren, D. (1997). Differences in vulnerability and susceptibility to dexamphetamine in Nijmegen high and low responders to novelty: A dose-effect analysis of spatio-temporal programming of behavior. *Psychopharmacology, 132,* 181–187.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999a, June 4). Genetics of mouse behavior: Interactions with lab environment. *Science, 284,* 1670–1672.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999b). *Multi-center trial of a standardized battery of tests of mouse behavior.* Retrieved from http://www.albany.edu/psy/obssr/

Crawley, J. N. (2000). *What's wrong with my mouse? Behavioral phenotyping of transgenic and knockout mice.* New York: Wiley-Liss.

Crawley, J. N., Belknap, J. K., Collins, A., Crabbe, J. C., Frankel, W., Henderson, N., et al. (1997). Behavioral phenotypes of inbred mouse strains. *Psychopharmacology, 132,* 107–124.

Drai, D., Benjamini, Y., & Golani, I. (2000). Statistical discrimination of natural modes of motion in rat exploratory behavior. *Journal of Neuroscience Methods, 96,* 119–131.

Drai, D., & Golani, I. (2001). SEE: A tool for the visualization and analysis of rodent exploratory behavior. *Neuroscience and Biobehavioral Reviews, 25,* 409–426.

Drai, D., Kafkafi, N., Benjamini, Y., Elmer, G. I., & Golani, I. (2001). Rats and mice share common ethologically relevant parameters of exploratory behavior. *Behavioural Brain Research, 125,* 133–140.

Eilam, D., & Golani, I. (1989). Home base behavior of rats (*Rattus norvegicus*) exploring a novel environment. *Behavioural Brain Research, 34,* 199–211.

Eilam, D., Golani, I., & Szechtman, H. (1989). D_2 agonist quinpirole induces perseveration of routes and hyperactivity but no perseveration of movements. *Brain Research, 460,* 255–267.

Elmer, G. I., Gorelick, D. A., Goldberg, S. R., & Rothman R. B. (1996). Acute sensitivity vs. context-specific sensitization to cocaine as a function of genotype. *Pharmacology Biochemistry and Behavior, 53,* 623–628.

Flint, J., Corley, R., DeFries, J. C., Fulker, D. W., Gray, J. A., Miller, S., & Collins, A. C. (1995, September 8). A simple genetic basis for a complex psychological trait in laboratory mice. *Science, 269,* 1432–1435.

Gingras, M. A., & Cools, A. R. (1997). Different behavioral effects of daily or intermittent dexamphetamine administration in Nijmegen high and low responders. *Psychopharmacology, 132,* 188–194.

Golani, I., Benjamini, Y., & Eilam, D. (1993). Stopping behavior: Constraints on exploration in rats (*Rattus norvegicus*). *Behavioural Brain Research, 53,* 21–33.

Gorris, L. G. M., & van Abeelen, J. H. F. (1981). Behavioral effect of (-)naloxone in mice from four inbred strains. *Psychopharmacology, 74,* 355–359.

Hatcher, J. P., Jones, D. N. C., Rogers, D. C., Hatcher, P. D., Reavill, C., Hagan, J. J., & Hunter, A. J. (2001) Development of SHIRPA to characterize the phenotype of gene-targeted mice. *Behavioural Brain Research, 125,* 43–47.

Hen, I., Sakov, A., Kafkafi, N., Golani, I., & Benjamini, Y. (2003). *The dynamics of spatial behavior: How can robust smoothing techniques help?* Manuscript submitted for publication.

Jones, B. C., Reed, C. L., Radcliffe, R. A., & Erwein, G. (1993). Pharmacogenetics of cocaine: I. Locomotor activity and self-selection. *Pharmacogenetics, 3,* 182–188.

Kafkafi, N. (in press). Extending SEE for large-scale phenotyping of mouse open-field behavior. *Behavioral Research Methods, Instruments and Computers.*

Kafkafi, N., Mayo, C. L., Drai, D., Golani, D., & Elmer, G. I. (2001). Natural segmentation of the locomotor behavior of drug-induced rats in a photobeam cage. *Journal of Neuroscience Methods, 109,* 111–121.

Kafkafi, N., Pagis, M., Lipkind, D., Mayo, C. L., Benjamini, Y., Elmer, G. I., & Golani, D. (in press). Darting behavior: A quantitative movement pattern for discrimination and replicability in mouse locomotor behavior. *Behavioural Brain Research.*

Logue, S. F., Owen, E. H., Rasmussen, D. L., & Wehner, J. M. (1997). Assessment of locomotor activity, acoustic and tactile startle, and prepulse inhibition of startle in inbred mouse strains and F1 hybrids: Implications of genetic background for single gene and quantitative trait loci analyses. *Neuroscience, 80,* 1075–1086.

Nolan, P. M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., et al. (2000). A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genetics, 25,* 440–443.

Paigen, K., & Eppig, J. T. (2000). A mouse phenome project. *Mammalian Genome, 11,* 715–717.

Paulus, M. P., & Geyer, M. A. (1993). Three independent factors characterize spontaneous rat motor activity. *Behavioural Brain Research, 53,* 11–20.

Rocha, B. A., Odom, L. A., Barron, B. A., Ator, R., Wild, S. A., & Forster, M. J. (1998). Differential responsiveness to cocaine in C57BL/6J and DBA/2J mice. *Psychopharmacology, 138,* 82–88.

Rogers, D. C., Jones, D. N., Nelson, P. R., Jones, C. M., Quilter, C. A., Robinson, T. L., & Hagan, J. J. (1999). Use of SHIRPA and discriminant analysis to characterise marked differences in the behavioural phenotype of six inbred mouse strains. *Behavioural Brain Research, 105,* 207–217.

Spink, A. J., Tegelenbosch, R. A. J., Buma, M. O. S., & Noldus, L. P. J. J. (2001). The EthoVision video tracking system—a tool for behavioral phenotyping of transgenic mice. *Physiology & Behavior, 73,* 731–734.

Szechtman, H., Culver, K., & Eilam, D. (1999). Role of dopamine systems in obsessive-compulsive disorder (OCD): Implications from a novel psychostimulant-induced animal model. *Polish Journal of Pharmacology, 51,* 55–61.

Takahashi, J. S., Pinto, L. H., & Vitaterna, M. H. (1994, June 17). Forward and reverse genetic approaches to behavior in the mouse. *Science, 264,* 1724–1733.

Tchernichovski, O., Benjamini, Y., & Golani, I. (1996). Constraints and the

emergence of freedom in the ontogeny of rat exploratory behavior. *Behaviour, 133,* 519–539.

Tchernichovski, O., Benjamini, Y., & Golani, I. (1998). The dynamics of long term exploratory behavior in the rat, Part I. *Biological Cybernetics, 78,* 423–432.

Tchernichovski, O., & Golani, I. (1995). A phase plane representation of rat exploratory behavior. *Journal of Neuroscience Methods, 62,* 21–27.

Tirelli, E., & Witkin, J. M. (1994). Verticalization of behavior elicited by dopaminergic mobilization is qualitatively different between C57BL/6J and DBA/2J mice. *Psychopharmacology, 116,* 191–200.

Tolliver, B. K., & Carney, J. M. (1995). Locomotor stimulant effects of cocaine and novel cocaine analogs in DBA/2J and C57BL/6J inbred mice. *Pharmacology Biochemistry and Behavior, 50,* 163–169.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Turri, M. G., Talbot, C. J., Radcliffe, R. A., Wehner, J. M., & Flint, J. (1999). High-resolution mapping of quantitative trait loci for emotionality in selected strains of mice. *Mammalian Genome, 11,* 1098–1101.

van der Staay, F. J., & Steckler, T. (2002). The fallacy of behavioral phenotyping without standardization. *Genes, Brain and Behavior, 1,* 9–13.

Wahlsten, D. (2001). Standardizing tests of mouse behavior: Reasons, recommendations and realty. *Physiology & Behavior, 73,* 695–704.

Wahlsten, D., Crabbe, J. C., & Dudek, B. C. (2001). Behavioral testing of standard inbred and 5HT1B knockout mice: Implications of absent corpus callosum. *Behavioural Brain Research, 125,* 23–32.

Wallace, D. G., Hines, D. J., & Whishaw, I. Q. (2002). Quantification of a single exploratory trip reveals hippocampal formation mediated dead reckoning. *Journal of Neuroscience Methods, 30,* 131–145.

Whishaw, I. Q., Cassel, J. C., Majchrzak, M., Cassel, S., & Will, B. (1994). "Short-stops" in rats with fimbria-fornix lesions: Evidence for change in the mobility gradient. *Hippocampus, 4,* 577–582.

Whishaw, I. Q., Hines, D. J., & Wallace, D. G. (2001). Dead reckoning (path integration) requires the hippocampal formation: Evidence from spontaneous exploration and spatial learning tasks in light (allothetic) and dark (idiothetic) tests. *Behavioural Brain Research, 127,* 49–69.

Womer, D. E., Jones, B. C., & Erwin, V. G. (1994). Characterization of dopamine transporter and locomotor effects of cocaine, GBR 12909, epidepride, and SCH 23390 in C57BL and DBA mice. *Pharmacology Biochemistry and Behavior, 48,* 327–335.

Wurbel, H. (2000) Behaviour and the standardization fallacy. *Nature Genetics, 26,* 263.

Wurbel, H. (2002). Behavioral phenotyping enhanced—beyond (environmental) standardization. *Genes, Brain and Behavior, 1,* 3–8.

Received June 19, 2002
Revision received October 15, 2002
Accepted November 12, 2002 ■

ORDER FORM

Start my 2003 subscription to *Behavioral Neuroscience*!

ISSN: 0735-7044

___ $121.00, APA Member/Affiliate _____
___ $213.00, Individual Nonmember _____
___ $563.00, Institution _____
In DC add 5.75% / In MD add 5% sales tax _____
Total Amount Enclosed $ _____

Subscription orders must be prepaid. (Subscriptions are on a calendar year basis only.) Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.

SEND THIS ORDER FORM TO:
American Psychological Association
Subscriptions
750 First Street, NE
Washington, DC 20002-4242

American Psychological Association

Or call (800) 374-2721, fax (202) 336-5568.
TDD/TTY (202) 336-6123.
For subscription information, e-mail:
subscriptions@apa.org

◯ Send me a FREE Sample Issue
◯ Check enclosed (make payable to APA)

Charge my: ◯ VISA ◯ MasterCard ◯ American Express

Cardholder Name _____
Card No. _____ Exp. Date _____
_____
Signature (Required for Charge)

BILLING ADDRESS: _____
_____
City _____ State _____ Zip _____
Daytime Phone _____

SHIP TO:
Name _____
Address _____
_____
City _____ State _____ Zip _____
APA Member # _____ *BNEA13*