NeuroToxicology® 18(4): 947-968, 1997

The IPCS Collaborative Study On Neurobehavioral Screening Methods: IV. Control Data

VIRGINIA C. MOSER[1§†], GEORGE C. BECKING[2§], VINCENZO CUOMO[3§†], EMIL FRANTÍK[4§†], BEVERLY M. KULIG[5§†], ROBERT C. MACPHAIL[1§], HUGH A. TILSON[1§], GERHARD WINNEKE[6§], W. STEPHEN BRIGHTWELL[7†], MARIA A. DE SALVIA[3†], MICHAEL W. GILL[8†], GILLIAN C. HAGGERTY[9†], MIROSLAVA HORNYCHOVÁ[4†], JAN LAMMERS[5†], JENS-JØRGEN LARSEN[10†], KATHERINE L. MCDANIEL[1†], B.K. NELSON[7†] AND GRETE ØSTERGAARD[10†]

§ Members of the Steering Group for the IPCS-sponsored Collaborative Study on Neurobehavioral Screening Methods. † Principal investigators and key personnel from the participating laboratories.

[1]Neurotoxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, U.S.A.; [2]International Programme on Chemical Safety, World Health Organization, Research Triangle Park, North Carolina, U.S.A.; [3]Institute of Pharmacology, University of Bari, Bari, Italy; [4]National Institute of Public Health, Prague, Czech Republic; [5]TNO Nutrition and Food Research, Zeist, The Netherlands; [6]Institute of Environmental Hygiene, Dusseldorf, Germany; [7]National Institute for Occupational Safety and Health, Cincinnati, Ohio, U.S.A; [8]Bushy Run Research Center, Union Carbide Corp., Export, Pennsylvania, U.S.A.; current address: Toxicology Regulatory Service, Inc., Charlottesville, Virginia, U.S.A.; [9]G.D. Searle and Co., Skokie, Illinois, U.S.A.; [10]Ministry of Health, National Food Agency, Soborg, Denmark.

Abstract: *The goal of the International Programme on Chemical Safety (IPCS) Collaborative Study on Neurobehavioral Screening Methods was to determine the intra- and inter-laboratory reliability of a functional observational battery (FOB) and an automated assessment of motor activity in eight laboratories worldwide. The control data were crucial to the outcome of the studies in terms of sensitivity and reliability of the test measures, which in turn impact on the between-laboratory comparisons of chemical effects. In addition, analyses of control data can aid in determining endpoints that may require modification to improve their sensitivity and reliability. The control data from the eight laboratories were examined in terms of the following parameters: 1) control variability within studies for each laboratory; 2) within-laboratory replicability of control values across studies; 3) within-laboratory stability of control values over the course of testing for a given study; and 4) between-laboratory comparisons of parameters (1), (2), and (3). The analyses indicated considerable differences across endpoints, wherein some measures showed high variability and little replicability, while others were extremely reproducible. Generally, there were similar ranges of variability and replicability of control data across laboratories, although in some cases one or two laboratories were markedly different from the others. The physiological (weight, body temperature) and neuromuscular (grip strength, landing foot splay) endpoints exhibited the least variability, whereas the subjective assessments of reactivity varied the most. These data indicate a reasonable degree of comparability in the data generated in the participating laboratories. ©1997 Intox Press, Inc.*

Keywords: Functional Observational Battery, Motor Activity, Between-laboratory Comparisons, Within-laboratory Comparisons, Control Data

INTRODUCTION

The International Programme on Chemical Safety (IPCS) sponsored a collaborative study to evaluate the utility of neurobehavioral test methods for screening neurotoxic chemicals (see Moser *et al.*, 1997a). The data compiled in this Collaborative Study provide ample opportunity to assess the reproducibility of control data, both between and within laboratories, on the various test measures. In any study, the control data are one of the crucial factors in terms of sensitivity of the test measure, i.e., what magnitude of change can be detected statistically

Please send requests for reprints to: Dr. V.C. Moser, NTD, (MD-74B), U.S. EPA, Research Triangle Park, NC 27711.

This paper has been reviewed by the National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation of use.

with a given number of subjects (power), and the reliability, i.e., how well the control data replicate across experiments. In turn, these factors impact on the between-laboratory comparisons of chemical effects. Another reason for scrutinizing control data is to determine which endpoints are robust, and which are candidates for modifications in procedure or technique.

According to the protocol (Moser *et al.*, 1997a), baseline (time-0) values were collected on each endpoint before dosing began. Since the rats had not been treated, all rats in each study were combined to provide baseline control data. All laboratories used five doses with either eight or ten rats/dose, yielding baseline control data on 40 or 50 rats from each study. Formal data collection was conducted using seven chemicals, under two dosing conditions each (acute and four-week repeated-dosing exposures), for a total of 14 studies from each laboratory. In this manuscript, baseline control data from 102 valid studies are presented (14 studies from each of eight laboratories, with some data excluded; see Moser *et al.*, 1997b).

The variability of these control data was evaluated, for each laboratory and each individual study, by examining the range of values in each sample comprising the baseline control mean. Replicability of these control data for each laboratory was defined as the variation of the baseline means across studies. Thus, control variability refers to the distribution of baseline values for any given study in a given laboratory, which impacts the sensitivity of the endpoint, whereas replicability refers to the expectation of obtaining similar mean values across studies in a given laboratory, which assesses the robustness of the endpoint. Finally, these same parameters are compared across laboratories to address questions of reproducibility between laboratories.

MATERIALS AND METHODS

Statistical Approach

The statistical approaches taken in this manuscript to describe control data are defined in Table 1. For continuous-variable measures, group means and standard deviations are calculated for each study from each laboratory. At time-0, the mean and coefficient of variation (C.V.: the standard deviation expressed as a percentage of the mean) for each sample are used to evaluate the variability with in each specific group of rats. For each laboratory, control variability is assessed by examining the baseline means and C.V.s from each study. For example, in a laboratory using 50 rats in each study, there would be a baseline mean and C.V. for each set of 50 rats from 14 studies, i.e., 14 baseline means and 14 baseline C.V.s. Replicability is represented by the grand mean, or the mean and range of baseline means, as well as the variability of baseline means, expressed as the grand C.V. In the above example, the grand mean and C.V. would be calculated as the average and the C.V. of the 14 baseline means; each laboratory would have one grand mean and grand C.V. for each endpoint. Thus, replicability indicates the expectation of repeatedly obtaining the same group data in different experiments. Finally, the grand means and grand C.V.s from all eight laboratories are averaged to produce an overall mean and an overall C.V.; note that the overall C.V. is the mean of grand C.V.s, not the C.V. of the grand means. Reproducibility is evaluated as the range of means and replicability in different laboratories.

For ordinal, or ranked endpoints, the distributions of scores are used to define control data for each study in each laboratory. These data are non-parametric, i.e., not normally distributed. Therefore, frequency distributions and medians of the baseline data are determined for each sample at time-0, but there is no measure of variability. To evaluate replicability, the baseline frequency distributions are averaged across all studies, providing a grand distribution which represents the mean and variability of frequencies associated with each possible score. The C.V.s of these frequencies are calculated, but for very low frequencies the C.V. either can not be calculated (because the mean is 0) or else is extremely large (as the mean approaches 0, the C.V. becomes inflated). Therefore for these minor peaks, arbitrarily defined as having frequencies $\leq$10.0%, the range of frequencies are listed instead of a C.V. The grand C.V. of the remaining peaks, having frequencies >10.0%, is taken as an assessment of replicability. Similarly, overall means and overall C.V.s of the grand distributions, evaluated across laboratories, are taken to represent reproducibility.

According to the protocol (Moser *et al.*, 1997a), each study included a concurrent, vehicle-treated control group. Therefore for each study, data exist for the control group (8 or 10 control rats/study) at each time point (four test times per study). In the acute studies, the rats were tested before dosing, at the time of peak effect (TOPE) on the day of dosing, 24 hours, and one week after dosing.

TABLE 1. Definitions and Descriptions of Terms Used to Evaluate Inter- and Intra-laboratory Stability of Control Data.

	Continuous Data	Discontinuous (ordinal) Data
Within-Laboratory		
Control Variability	mean, standard deviation of time-0 samples, n=40-50 rats	frequency distribution of time-0 samples, n=40-50 rats
	Baseline mean: mean of each time-0 sample	Baseline median: median of time-0 distributions
	Baseline C.V.: C.V.s of each time-0 sample	—
Replicability	mean, standard deviation of baseline means, n=14 studies	mean, standard deviation of each frequency of each possible score, n=14 studies
	Grand mean: mean of baseline means	Grand distribution: mean frequencies of each score
	Grand C.V.: C.V. of baseline means	C.V.: C.V. of mean frequencies of each score
		Grand C.V.: mean of C.V.s of peaks with distribution frequencies >10% of sample
Between-Laboratory		
Reproducibility	mean, standard deviation of grand means and C.V.s, n=8 laboratories	mean, standard deviation of grand distributions and C.V.s, n=8 laboratories
	Overall mean: mean of grand means	Overall mean: mean of grand distributions of major peaks
	Overall C.V.: mean of grand C.V.s	Overall C.V.: mean of grand C.V.s

In the repeated-dosing studies, tests were conducted before dosing, during the second and fourth weeks of dosing, and two weeks after dosing ended. These data provide information on how the control values changed with repeated testing over time. As described above, continuous data are represented by means and variability estimates, whereas ranked data are expressed as distributions of scores.

RESULTS

Baseline control data from all valid studies are presented in this manuscript. There was a total of 14 studies for most laboratories, with the exception of laboratory G (with 13 "valid" studies) and laboratory C (five "valid" studies; see Moser *et al.*, 1997b). The individual tests of the screening battery were grouped into several domains of neurobiological function (see Moser *et al.*, 1997a). The data for each endpoint are described below, presented by functional domain.

Activity Measures

The measures comprising the Activity domain included motor activity, rearing, and home-cage posture. The data for home-cage posture showed almost no variation for control rats, i.e., all rats were described as displaying normal postural activities: sitting, standing, or rearing. Data for rearing and motor activity showed considerably more diversity.

Rearing. The baseline values for the number of rears in the open field, across all studies, are presented in Figure 1 and summarized in Table 2. There was considerable control variability within each study, expressed by the C.V.s which averaged 31% to 59% of the group means. However, the actual group means and range of data were quite similar both within and between laboratories. For each laboratory, the means from all studies were averaged, and the grand means are presented in Figure 2 and Table 2. There was generally good replicability of the group mean data, with grand C.V.s generally ≤15% for six of the eight laboratories. Across laboratories, the overall mean was 10 rears, with an overall C.V. of 15%

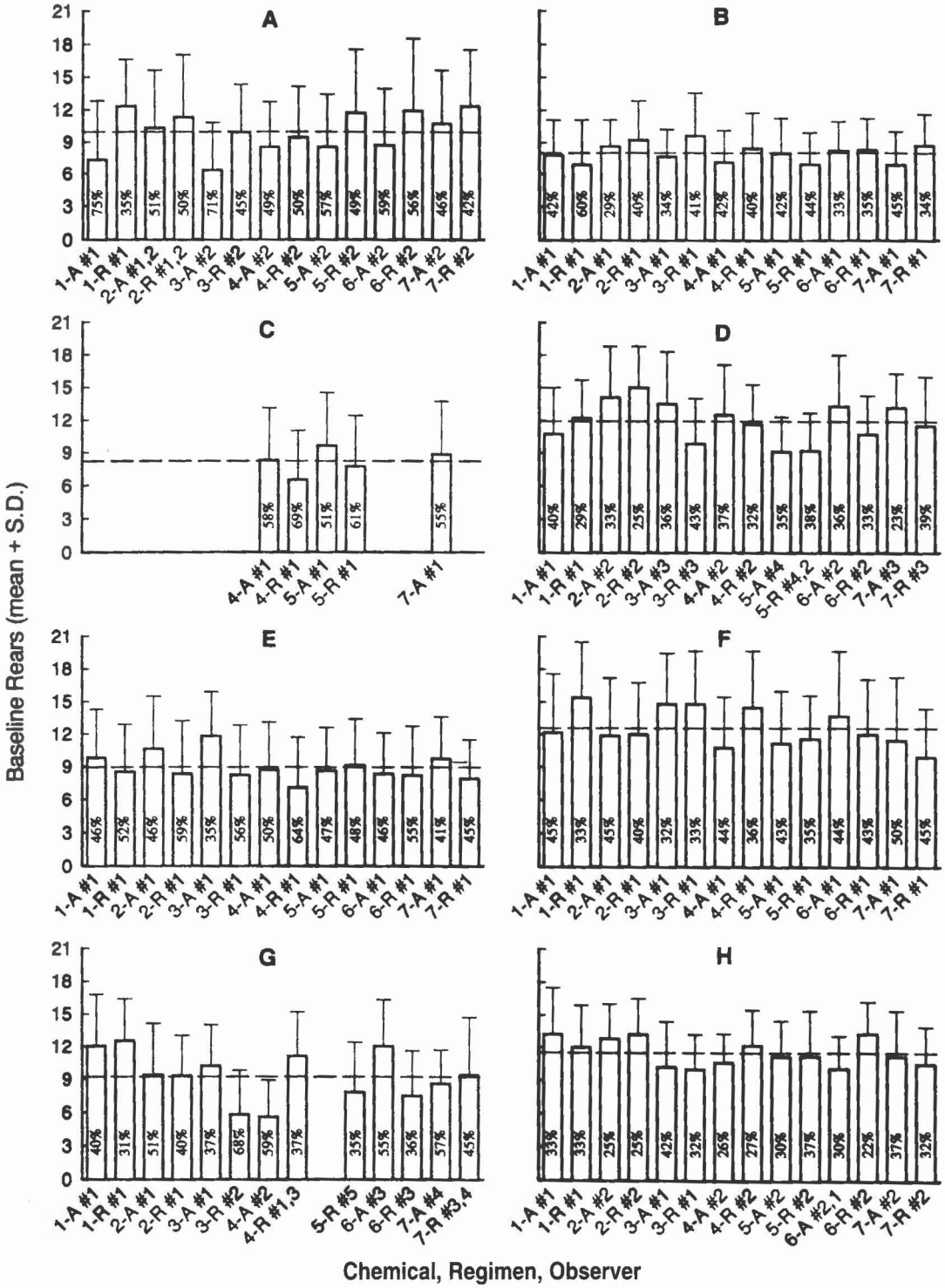

FIG. 1. Baseline control data for the number of rears in the open field. For each laboratory (A-H), the baseline mean + standard deviation of the time-0 samples (n=40 or 50 rats/study) are plotted, with the baseline C.V. listed within each bar. The independent axis indicates the chemical number (#1-7), dosing regimen (A=acute, R=repeated), and the observer number (where two numbers are listed for one study, both observers were involved in testing the rats). The dashed line indicates the grand mean (mean of the control means) across studies.

(listed in Table 7). Thus, despite considerable variability within groups of rats, similar group means were obtained within the same laboratory, and between laboratories.

Evaluation of the control groups across the studies showed that within each laboratory, the mean values generally decreased with repeated testing (data not shown). This was most obvious comparing the time-0 test (means ranging from 6-15) and the second test (means ranging from 2-8). The subsequent changes between the second and third, and the third and fourth, tests were not as great. Variability also greatly increased with repeated testing and C.V.s sometimes exceeded 100%. The reason for this was that the rate of habituation varied greatly between rats in the same group: whereas at time-0 almost all rats rear at least several times, with repeated testing some rats no longer rear at all while others continue to show baseline numbers of rears. Such large variability implies that treatment-related changes of smaller magnitude would not be statistically significant; however, the square-root transformation routinely performed on the data before analysis (to normalize discrete distributions, see Moser *et al.*, 1997a) markedly lowered the variability of the data.

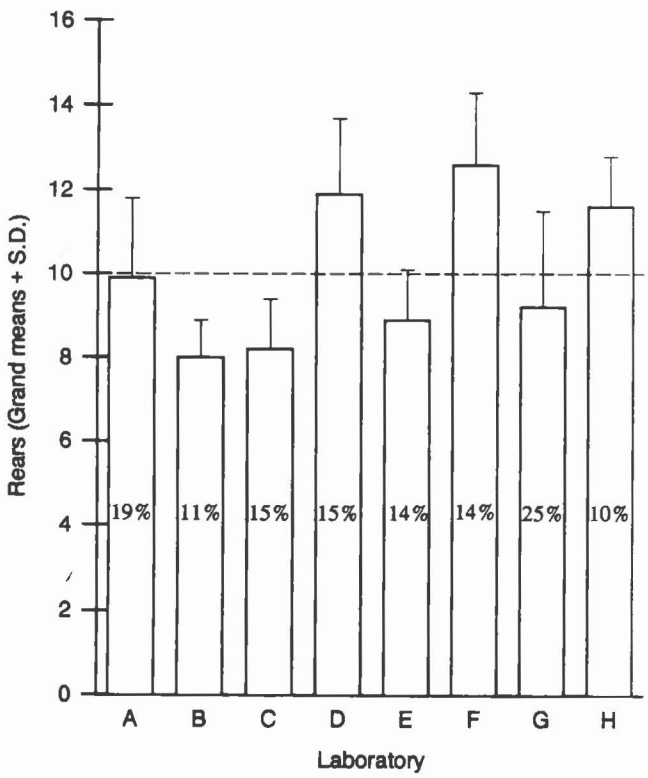

FIG. 2. Grand control means for the number of rears in the open field. For each laboratory (A-H), the grand mean (indicated as dashed line in Figure 1) + standard deviation of the baseline means are presented (n=14 studies, except for laboratories C and G), with the grand C.V. listed within each bar. The dashed line indicates the overall mean (mean of the grand means) across laboratories.

Motor Activity. Baseline values for motor activity data (total activity units during the test session) are also presented in Table 2. There were substantial differences in the actual values, partly due to the dependent measure recorded (e.g., distance traveled vs. photocell interruptions) and to the different activity devices used (see Moser *et al.*, 1997a). Interestingly, in laboratories which used essentially the same device (for example, laboratories B and E), the baseline data differed. The mean activity values across studies were generally stable for most laboratories, but three laboratories (B, F, G) showed two- to five-fold differences in group means even though the testing conditions remained constant across studies. This is also evident in the grand C.V., which was 6-15% in five laboratories but was 20-34% in those three. There were also pronounced differences in individual laboratory control variability. Two laboratories reported an average variability of 20% and 22%, and in those laboratories, the C.V. never exceeded 26% for any individual study. Others reported greater control variability (32-53%), which was as high as 78% and 116% in individual studies.

Due to the different units for which activity was recorded, an overall mean could not be calculated. The overall C.V. of 16% actually ranged from 6-34%. Thus, as was the case with rears, there was considerable variability within groups of rats but the group means were generally less variable; however, the laboratories showed marked individual differences in the variability of this measure.

There were also considerable differences in the pattern of control activity data across time. To illustrate this, the activity data for control groups at each test time for acute studies are presented in Figure 3, and for repeated-dose studies in Figure 4. For acute studies, where testing took place over a shorter period of time, the control values decreased mostly at the time of peak effect and at 24 hours. The magnitude of this decrease varied across laboratories, with one laboratory (C) averaging a 50% drop in activity levels after the first test. In the repeated-dose studies, most laboratories showed only a modest decline in activity, although one laboratory showed markedly decreased activity (laboratory C) while another obtained increasing activity (D). Most laboratories showed increasing variability as the control groups were repeatedly tested (data not shown). The C.V.s remained in the 30-50% range for three laboratories in the acute studies, and for five laboratories in the repeated-dose studies. In the remaining laboratories the range of C.V.s approached 40-60%, and there were instances in which the variability exceeded 100%.

TABLE 2. Baseline Control Data for the Number of Rears in the Open Field (during the observation period), and for Motor Activity in an Automated Device (total activity units during the session)[1].

Laboratory	Baseline means				C.V.s of baseline means[4]		
	Grand mean[2]	Grand C.V.	Range[3] low	Range[3] high	Mean	Range low	Range high
	REARS						
A	10	19%	6	12	53%	35%	75%
B	8	11	7	10	40	29	60
C	8	15	7	10	59	51	69
D	12	15	9	15	34	23	43
E	9	14	7	12	49	35	64
F	13	14	10	15	41	32	50
G	9	25	6	13	45	31	68
H	12	10	10	13	31	22	42
	MOTOR ACTIVITY						
A	1368	12%	1137	1607	35%	21%	78%
B	677	20	403	951	39	28	56
C	170	9	155	189	43	39	46
D	721	10	583	812	32	24	45
E	282	6	255	313	22	17	26
F	128	20	86	168	40	30	53
G	161	34	46	247	53	38	116
H	6127	15	4976	7677	20	17	26

[1]Data for 14 studies are represented for all laboratories except laboratory G (13 studies) and C (5 studies).
[2]Mean of time-0 or baseline means for each laboratory, and C.V. of the baseline means
[3]Range of group means at time-0 for each laboratory
[4]Mean and range of C.V.s at time-0 for each laboratory

Autonomic Measures

Lacrimation, Salivation, Pupil Response, Palpebral Closure. The data for lacrimation, salivation, pupil response, and palpebral closure were listed as 'normal' (e.g., no lacrimation; pupil response present) in most rats at time-0 and in control rats across time. At time-0, the pupil response was detected in ≥99% of rats by all but one laboratory (laboratory D: 97% of rats). The response was also evident in control rats (>99%) throughout the studies in six laboratories, but the response was recorded in fewer rats (89% and 95%) in two laboratories (B and D, respectively). Salivation and lacrimation were virtually never recorded, with one exception: laboratory B recorded a total of 17 instances of lacrimation in control rats across all studies.

Urination and Defecation. On the other hand, the assessment of fecal and urinary output displayed considerable between-laboratory differences (data not shown). Four laboratories recorded a high number of rats with no urination (42-64%), and progressively decreasing numbers of rats showing greater numbers of pools. Two other laboratories generally showed a flat distribution of the number of pools (10-25% each for the occurrence of one to five pools). Two laboratories reported no urination in most rats (43-63%), but there were also many instances of 'polyuria' (16-29%). The same pattern of control distributions was obtained for the number of fecal boluses, with five laboratories recording no defecation for the majority of rats (58-87%). The remaining laboratories showed a flatter distribution across bolus counts, and one of these reported a relatively high number (34%) of rats showing more than four boluses (sometimes as many as 8-11 boluses). Excretion distributions were not consistent across studies, but in general control rats usually had fewer excretions over the course of testing.

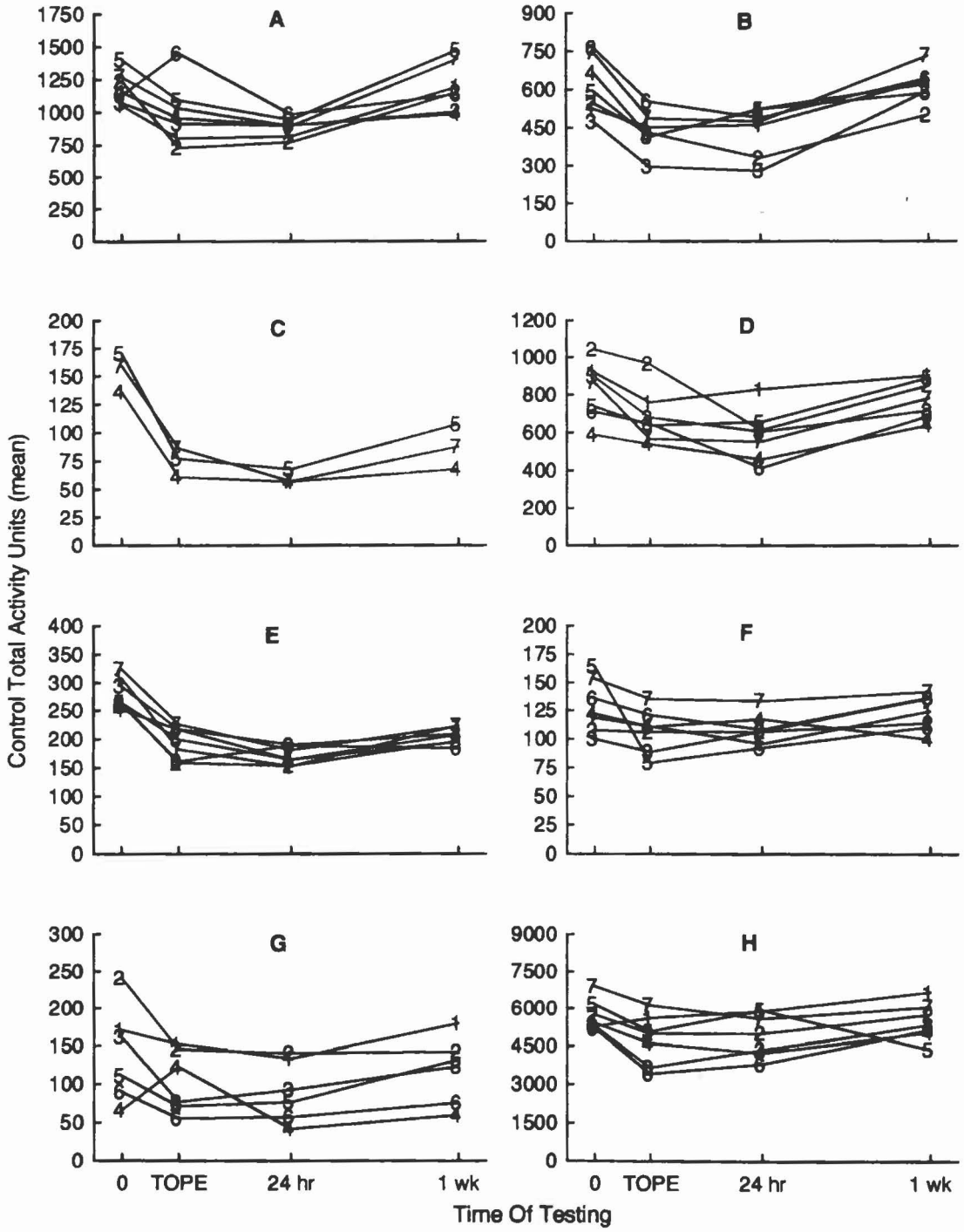

FIG. 3. Motor activity data across testing within each acute study. For each laboratory (A-H), the mean activity units for the control groups (n=8 or 10 rats/study) are plotted for each test time (time-0, time of peak effect (TOPE), 24 hours, and one week after dosing). The symbols of each line indicate the chemical numbers (#1-7), i.e., study number.

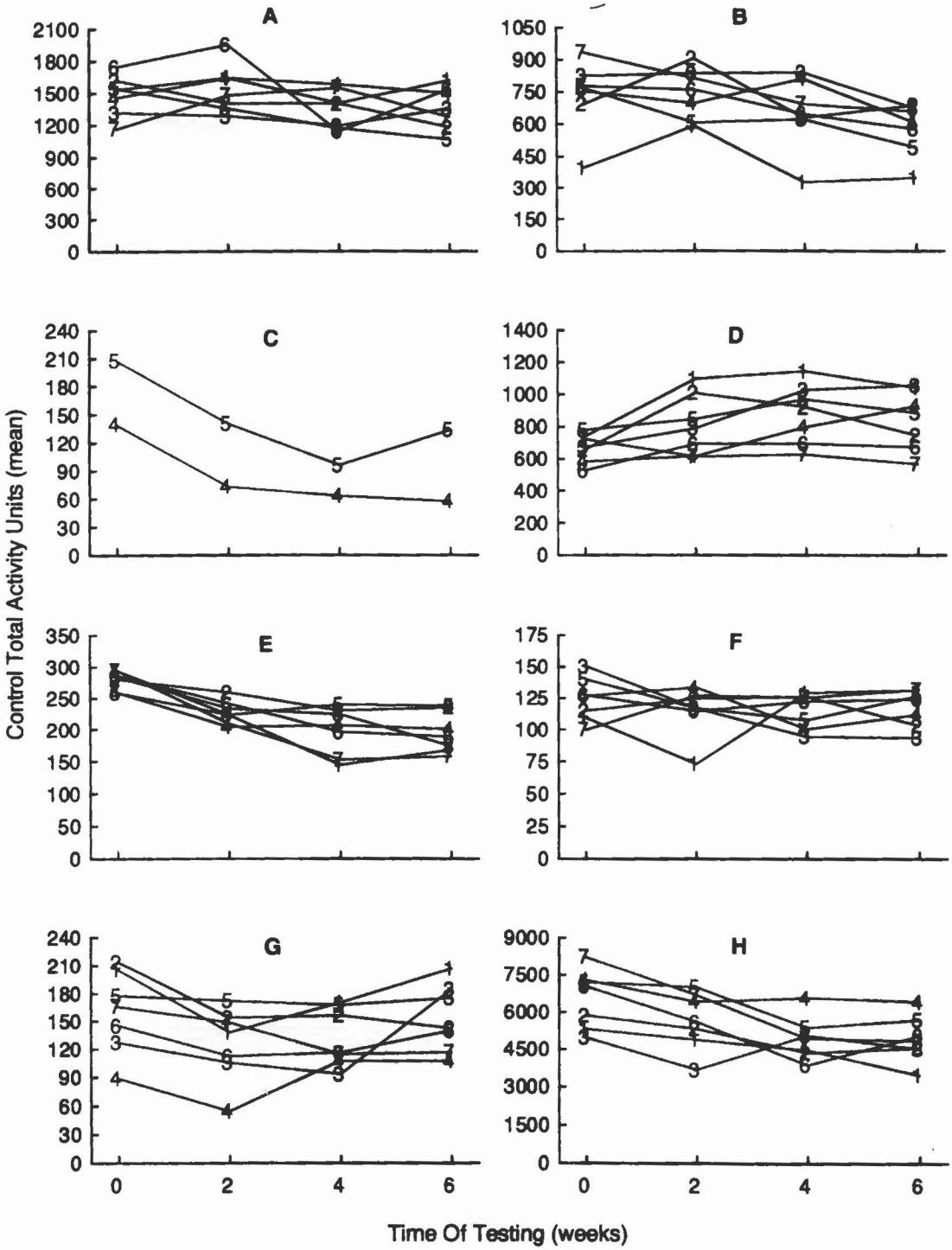

FIG. 4. Motor activity data across testing within each repeated-dose study. For each laboratory (A-H), the mean activity units for the control groups (n=8 or 10 rats/study) are plotted for each test time (time-0, weeks two and four during dosing, and week six, two weeks after dosing ended). The symbols of each line indicate the chemical numbers (#1-7), i.e., study number.

Excitability Measures

Handling Reactivity and Ease of Removal. The measures assessing excitability and general reactivity varied in reproducibility both between and within laboratories. Figure 5 depicts the baseline data for handling reactivity presented as frequency polygons, or distributions, from each study for all laboratories. The grand distributions, or mean frequencies of each score across studies, are shown in the graphs as a solid line (± one standard deviation), and are listed in Table 3. The distributions of baseline data for ease of removal were somewhat similar to those for handling reactivity, and are also listed in Table 3.

For both measures, six laboratories scored a majority of rats with '2's, but the percentage ranged from 92% of the rats (laboratory D) to around 50-60% (laboratory A, whose remaining scores were mostly '1's; and B, whose remaining scores were mostly '3's). Laboratory C scored primarily '1's and some '2's. In contrast, laboratory G scored on average more '3's than '2's; however, this distribution varied greatly across studies. Replicability of baseline data across studies can be assessed visually by examining Figure 5. In Table 3, variability of the distributions is listed as the C.V. of the frequencies of each score (e.g., the mean and C.V. of the percentages of rats receiving a '2' across all studies) and the grand, or average, C.V. of the larger peaks of the distribution (frequencies >10%). Compare, for example, the relatively constant distributions obtained in laboratory D with a grand C.V. of 8%, with the pronounced differences across studies in laboratory G whose grand C.V. was 67% (Figure 5).

TABLE 3. Baseline Control Data for the Excitability Endpoints: Arousal, Ease of Removal (both scored from 1 to 6), and Handling Reactivity (scored from 1 to 4)[1].

Laboratory	Mean percentage or range of rats receiving each score (C.V.)[2]						Baseline medians[3]	Grand C.V.[4]
	1	2	3	4	5	6		
	HANDLING REACTIVITY							
A	24 (62)	62 (24)	13 (77)	0-6			2	54%
B	0-50	52 (14)	37 (38)	0-8			1.5-3	26
C	76 (18)	24 (57)	0	0			1	38
D	0-6	92 (8)	0-20	0-8			2	8
E	10 (119)	80 (12)	0-18	0			2	66
F	0	65 (18)	25 (31)	2-20			2-3	25
G	0	43 (74)	47 (60)	0-88			2-4	67
H	0-26	75 (25)	17 (75)	0-29			2	50
	EASE OF REMOVAL							
A	26 (87)	51 (34)	15 (90)	0-16	0-12	0	1-3	70%
B	0-48	57 (22)	37 (38)	0-8	0-2	0-2	2-3	30
C	64 (10)	33 (26)	0-5	0	0-3	0	1	18
D	0-6	92 (11)	0-26	0-20	0-2	0	2	11
E	12 (102)	60 (20)	22 (36)	0-8	0-4	0-2	1.5-2.5	53
F	0-3	66 (14)	14 (42)	0-13	10 (45)	0-8	2	34
G	0	37 (81)	55 (57)	0-20	0-75	0	2-5	69
H	0-34	58 (20)	0-32	16 (57)	0-20	0-2	2-3	39
	AROUSAL							
A	0-2	0-2	0-6	69 (14)	28 (33)	0-6	4	24%
B	0	0	0-40	88 (12)	0-18	0	4	12
C	0	0	0-3	69 (7)	30 (16)	0	4	11
D	0	0	0-14	92 (11)	0-24	0	4	11
E	0	0	0-8	97 (3)	0-2	0	4	3
F	0	0	0-20	75 (20)	18 (78)	0	4	49
G	0	0	0-2	95 (6)	0-18	0-3	4	6
H	0	0	0-6	97 (2)	0-4	0	4	2

[1] Data for 14 studies are represented for all laboratories except laboratory G (13 studies) and C (5 studies)
[2] Mean and C.V. of frequency distributions of the baseline control means, where the frequency >10.0%; otherwise range of frequencies are listed
[3] Range of baseline medians across studies
[4] Mean of C.V.s of the peaks of the baseline distribution where the frequency >10.0%

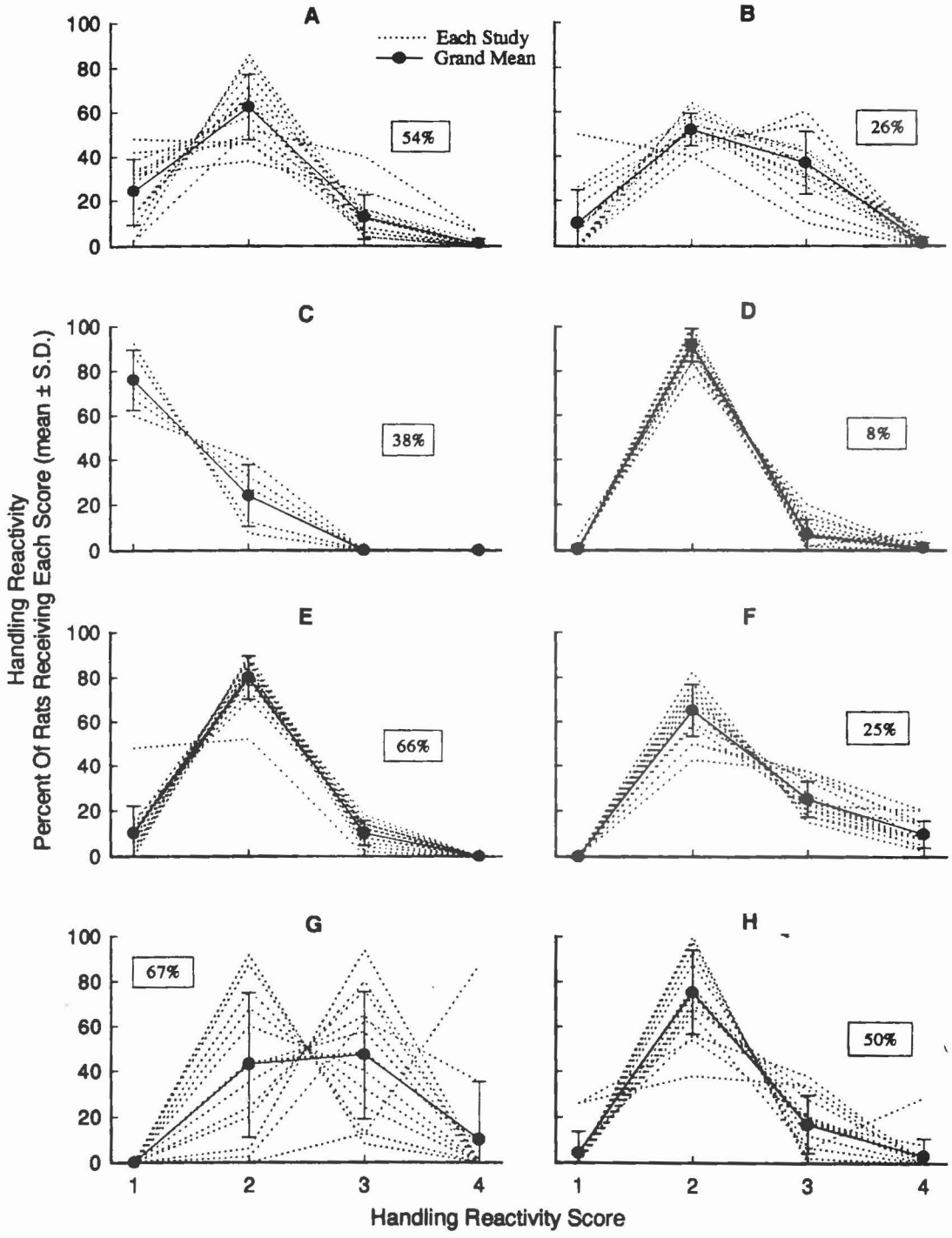

FIG. 5. Frequency polygons of the distributions of scores for handling reactivity, i.e., the percentage of rats receiving each of the possible scores (1 to 4). For each laboratory (A-H), the frequency polygon for the time-0 samples (n=40 or 50 rats/study) are plotted (dashed lines). The grand distributions, or mean ± standard deviation of the frequencies of each score across studies are also plotted (solid line). The grand C.V., i.e., mean of C.V.s of the major peaks of the grand distribution, are indicated for each laboratory.

In six laboratories (all but laboratories G and A), the profiles obtained with handling reactivity mirrored those for the ease of removal assessment (see Table 3). Laboratory G recorded large differences in baseline values for both measures (medians for ease of removal ranged from 2 to 5, and for handling reactivity medians were 2-4), and this variability was evident both within and between observers. Laboratory A reported considerably more variable data across studies and observers for the ease of removal measure than for handling reactivity. Finally, the data from laboratory B indicated a trend of increasing reactivity values over the first three studies, but the data were more stable thereafter.

Control values across time for both the ease of removal and handling reactivity scores generally either stayed the same or showed no consistent pattern of change. Few laboratories showed any decreases with repeated testing, with the exception of laboratory E which regularly recorded lower scores, particularly during the course of the repeated-dose studies.

Arousal. The baseline data distributions (percentage of rats receiving each score) for the ranking of arousal in the open field are presented in Table 3. All laboratories scored mostly '4's at time-0, indicating a typical level of exploration in the novel environment. In five laboratories, an average of ≥88% of the rats received a score of '4'; in the remaining laboratories the proportion was 69-75% with most of the remaining rats receiving '5'. The distributions of the data across studies were also consistent for all laboratories. Replicability of the group data was generally very good, and grand C.V.s were ≤12% in six laboratories. Control groups over the course of testing generally showed decreases in arousal due to habituation to the open field, and the median group scores approached '3' (data not shown). This pattern was clearly evident in six laboratories, while the remaining laboratories either showed no clear pattern of change (i.e., the trend differed across studies) or else the scores remained about the same across time.

Neuromuscular Measures

Landing Foot Splay. Baseline control data for landing foot splay are presented in Table 4. Splay values were generally similar across laboratories (about 70-90 mm), although laboratory H reported somewhat smaller values (~50 mm) which probably reflected the smaller size of their rats. Control variability was low: about 15-20% across studies in seven laboratories, and individual studies also showed C.V.s <30%. The eighth laboratory displayed an average C.V. of 28%, ranging up to 38%. There was no apparent influence of observer on this measure. Baseline means typically varied by ≤10%, and up to 14% in one laboratory: thus these data were quite reproducible within all laboratories.

Across laboratories, the overall mean value was 78 mm, with a low overall C.V. of the group means of 8%, and an overall average control variability of 20% (Table 7). Variability of the data did not change appreciably during the course of testing, and for seven laboratories the C.V. of the control group rarely exceeded 20-30%; however, laboratory G obtained C.V.s between 30% and 42% in 11 studies. The control data were mostly constant across the course of study as well, with only a slight downward trend in the repeated-dosing studies (data not shown). Foot splay was apparently not related to body weight, since the values did not increase over time in spite of the increasing weights of the rats.

Forelimb Grip Strength. Forelimb grip strength values are also presented in Table 4, and baseline means for each study in each laboratory are illustrated in Figure 6. Values were similar (approximately 0.8-1.1 kg) across most laboratories, with the exception of laboratory A who consistently reported higher grip strengths (~1.4 kg). Average values within each laboratory were mostly constant from study to study as well, as evidenced by most grand C.V.s <12%; laboratory G, however, showed the widest range of mean values (grand C.V.=20%). There were only slight differences between observers. The control variability within baseline groups averaged from <5% to 17%. Laboratory C reported the most invariant test values across studies (grand C.V.<0.5%) and also showed the lowest variability within studies (mean C.V.=3%). During the course of study, control group means generally remained constant in spite of increasing body weight of the rats. One exception to this was evident in laboratory G, but only with one particular observer: grip strength values doubled during the course of that observer's studies. Variability of control groups remained almost always <30% throughout testing: in only one laboratory, during only one study, did the C.V. of the control group reach 41%.

Hindlimb Grip Strength. Baseline values for hindlimb grip strength are shown in Table 4, and as was the case with forelimb grip, values were similar across six laboratories (mean values of 0.6-0.8 kg) with laboratories A and C reporting higher values (1.1-1.2 kg). The baseline groups showed somewhat more control variability in all laboratories, with mean C.V.s from 4-20%; however, variability sometimes approached 30-40% in particular stud-

TABLE 4. Baseline Control Data for Landing Foot Splay (mm), Forelimb Grip Strength (kg), and Hindlimb Grip Strength (kg)[1].

Laboratory	Baseline means				C.V.s of baseline means[4]		
	Grand mean[2]	Grand C.V.	Range[3] low	Range[3] high	Mean	Range low	Range high
	LANDING FOOT SPLAY						
A	92	9%	80	107	18%	14%	21%
B	87	7	74	96	17	14	20
C	91	2	84	93	21	19	23
D	86	9	77	98	18	14	22
E	77	10	60	89	21	16	26
F	72	6	65	81	20	17	26
G	70	14	52	86	28	22	38
H	52	4	48	57	14	11	19
	FORELIMB GRIP STRENGTH						
A	1.37	7%	1.21	1.49	8%	2%	12%
B	0.92	9	0.76	1.02	15	12	18
C	1.04	0[a]	1.04	1.05	3	2	4
D	0.80	11	0.64	0.91	17	13	22
E	1.00	9	0.84	1.10	14	10	20
F	0.90	12	0.75	1.12	17	13	23
G	1.06	20	0.70	1.48	15	9	28
H	0.92	7	0.80	1.00	13	11	17
	HINDLIMB GRIP STRENGTH						
A	1.07	11%	0.82	1.22	10%	3%	18%
B	0.78	11	0.62	0.91	20	17	33
C	1.20	1	1.18	1.23	4	4	6
D	0.61	16	0.45	0.76	15	11	19
E	0.80	9	0.67	0.93	20	17	24
F	0.78	21	0.52	1.02	18	14	24
G	0.81	27	0.32	1.27	18	11	41
H	0.59	7	0.54	0.68	12	9	17

[1]Data for 14 studies are represented for all laboratories except laboratory G (13 studies) and C (5 studies).
[2]Mean of time-0 or baseline means for each laboratory, and C.V. of the baseline means
[3]Range of group means at time-0 for each laboratory
[4]Mean and range of C.V.s at time-0 for each laboratory
[a]percentage < 0.5%

ies. While most laboratories reported relatively consistent hindlimb grip strength values across studies, with the grand C.V.s ≤11%, there were three exceptions: 1) laboratory G reported four-fold differences in control group values (0.3-1.3 kg), which could be attributed to different observers (i.e., certain observers always had higher or lower values); 2) laboratory D showed some differences (0.5-0.8 kg) which could not be reconciled with different observers, and 3) laboratory F showed an upward trend over time, with values doubling between the first and last study (0.5 and 1.0 kg, respectively).

Within each study, most laboratories obtained values that were relatively constant throughout a study. Control means across testing, averaged across all studies for each laboratory, are presented in Figure 7 to illustrate the overall trend of changes with repeated testing. As was the case with forelimb grip strength, laboratory G showed considerable increases across testing only with one particular observer: this was the reason for the increasing trend in grip strength values during the course of some repeated-dose studies for that laboratory (Figure 7). Variability of the control groups across studies also followed the same pattern observed with forelimb grip strength, with C.V.s <30% in most datasets.

Gait and Righting. Gait score was generally ranked as '1', indicating no abnormality at time-0. In a few laboratories, a small percentage of rats (<3%) were scored as having a slight ('2') to moderate ('3') abnormality. During

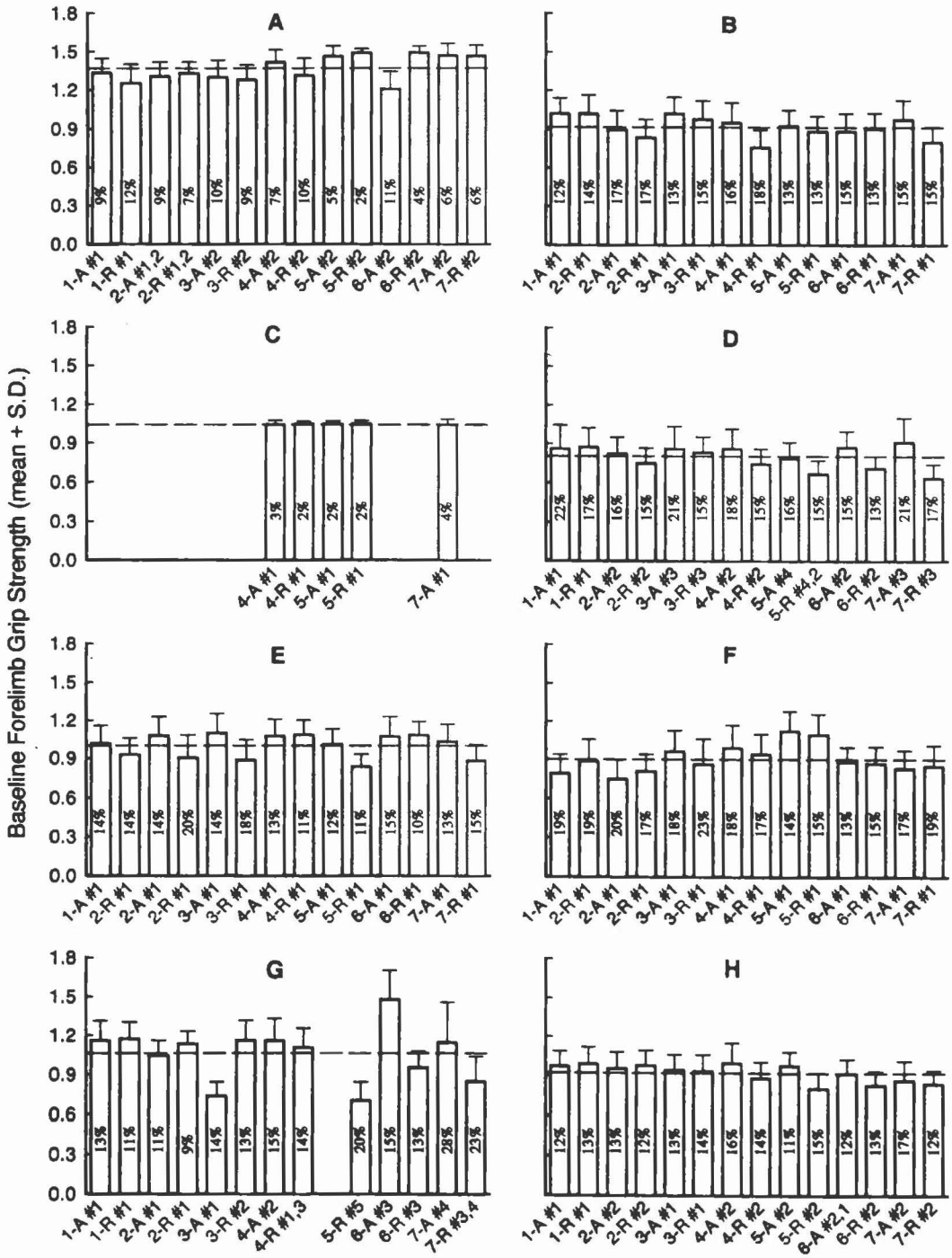

FIG. 6. Baseline control data for forelimb grip strength (kilograms to release). For each laboratory (A-H), the baseline mean + standard deviation of the time-0 samples (n=40 or 50 rats/study) are plotted, with the baseline C.V. listed within each bar. The independent axis indicates the chemical number (#1-7), dosing regimen (A=acute, R=repeated), and the observer number (where two numbers are listed for one study, both observers were involved in testing the rats). The dashed line indicates the grand mean (mean of the control means) across studies.

testing, control scores remained '1' in most laboratories, or else a few (one to four) rats would receive a '2'. In only a few specific studies, certain laboratories reported increasing percentages of '2's and '3's in controls across time. As with gait score, control scores for righting reflex were almost always '1' (no abnormality). Two laboratories (F, A) reported 3-6% '2's, and ~1% '3's in rats at time-0. These laboratories (and also B) reported a few altered righting scores in control rats (usually only one to two rats in the control group) during the course of the studies as well.

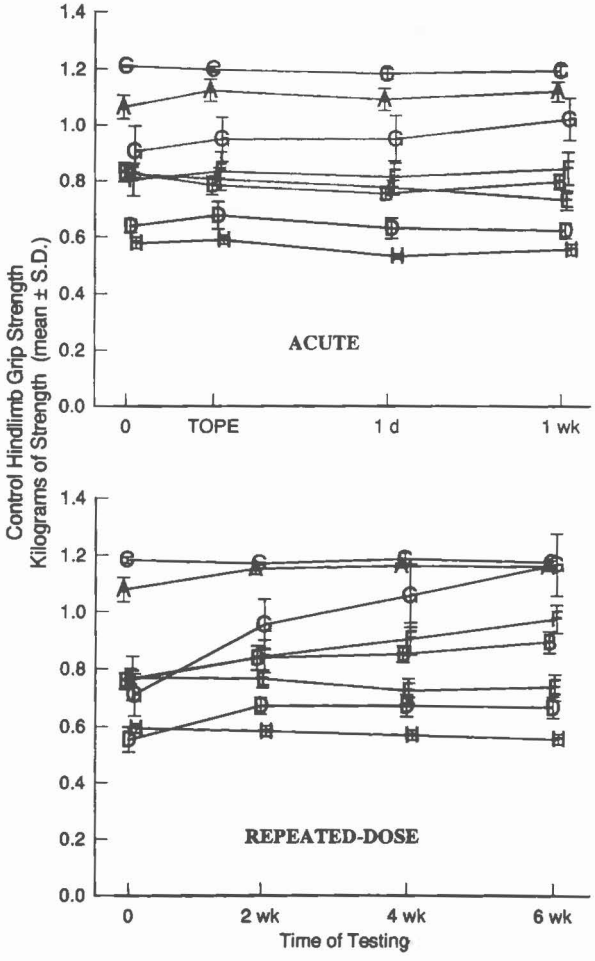

FIG. 7. Hindlimb grip strength data (kilograms to release) across testing within each study. For the acute studies (top), the means for the control groups (n=8 or 10 rats/study) at each time point (time-0, TOPE, 24 hours, one week) from all studies within each laboratory are averaged, and the mean ± standard deviation are plotted. Likewise for the repeated-dose studies (bottom), the means for the control groups (n=8 or 10 rats/study) at each time point (time-0, two, four, and six weeks) from all studies within each laboratory are averaged, and the mean ± standard deviation are plotted. The symbols of each line indicate the laboratory (A-H).

Sensorimotor Measures

Each of the sensorimotor responses in the FOB were ranked on the same scale ranging from '1' (no response) to '5' (exaggerated response; see Moser *et al.*, 1997a). Distributions of the baseline data are listed in Table 5.

Approach Response. The approach response was recorded mostly as very low reactivity, with the group medians almost always 2. Although more than half of the rats received a score of '2' in all laboratories, the percentage of rats with '2's ranged from 56-95%. The distribution of rankings was positively skewed in two laboratories (i.e., the remaining rats received mostly '3's), and negatively skewed in four laboratories (the remaining rats scored mostly '1's). The range of consistency in scoring is illustrated in the frequency polygons plotted in Figure 8. Two laboratories (A, F) recorded consistent scoring across studies, with the grand C.V.s of 5%. Two other laboratories (C, E) appeared somewhat less consistent, but the general pattern across studies was still similar; these laboratories had C.V.s of 25-28%. The remaining laboratories showed considerably less replicability, and grand C.V.s ranged from 47% to 64%. Differences between observers were not clearly evident in the pattern of scoring. Over the course of study, three laboratories recorded progressively lower scores in the control groups, whereas the scores remained about the same in four laboratories. The eighth laboratory did not show a clear pattern of change.

Touch Response. The distributions of baseline data for the touch response are also presented in Table 5. As was the case with approach response, most rats (48-89%) were scored as '2' but the distributions varied somewhat more across laboratories. The remaining rats received mostly scores of '1's and '3's, but unlike the approach response, three laboratories reported 10-20% of rats receiving '4's. Within each laboratory, the scoring of responses across studies also varied, with the grand C.V.s ranging from 8% to 73%. Control groups in most laboratories stayed fairly constant during testing; in only two laboratories were decreases in reactivity evident over testing with a study.

Click Response. In contrast to the consistently low approach and touch responses, the click response showed somewhat higher baseline scores and considerably more between-laboratory differences (Table 5). Median scores, both within and between laboratories, ranged from 2 to 4. The baseline distributions ranged from almost all '2's (e.g., laboratory B), to one with similar percentages of '2's and '3's (laboratory G), to almost all '3's (laboratory A), to a similar occurrence of '3's and '4's (laboratory H). Laboratories

TABLE 5. Baseline Control Data for the Sensorimotor Responses to an Approach, Touch, Click, and Tail-Pinch[1].

Laboratory	Mean percentage or range of rats receiving each score (C.V.)[2]					Baseline medians[3]	Grand C.V.[4]
	1	2	3	4	5		
	APPROACH RESPONSE						
A	0-10	95 (5)	0-4	0-4	0-2	2	5%
B	21 (86)	63 (27)	15 (79)	0-4	0-2	1-2	64
C	32 (31)	56 (20)	5-15	0	3-8	2	25
D	0-22	65 (31)	25 (88)	0-10	0	2-3	59
E	13 (50)	86 (7)	0-2	0-2	0-2	2	28
F	0-5	89 (5)	0-8	0-10	0-3	2	5
G	35 (78)	62 (44)	0	0-10	0	1-2	61
H	0-18	76 (19)	0-6	18 (75)	0	2	47
	TOUCH RESPONSE						
A	14 (60)	65 (35)	20 (125)	0-8	0	2-3	73%
B	0-40	87 (16)	0-25	0-6	0-2	2	16
C	13 (52)	81 (7)	0-18	0	0	2	29
D	0-20	58 (49)	34 (81)	0-30	0	2-3	65
E	2-14	89 (8)	0-14	0-6	0-2	2	8
F	0-3	88 (6)	0-3	10 (53)	0-3	2	29
G	0-26	73 (23)	0-15	12 (98)	0-3	2	61
H	0-10	48 (26)	29 (49)	20 (40)	0-2	2-3	38
	CLICK RESPONSE						
A	0	0-16	80 (17)	14 (73)	0-6	3	45%
B	0-6	87 (9)	0-8	0-28	0-2	2	9
C	0-8	13 (32)	63 (8)	19 (57)	0-13	3	32
D	0-2	26 (106)	64 (41)	0-52	0	2-4	74
E	0-2	25 (44)	62 (16)	13 (61)	0-4	3	40
F	0-3	56 (21)	29 (29)	14 (81)	0-3	2-3.5	44
G	0	40 (62)	46 (47)	14 (58)	0-3	2-3	56
H	0-2	0-12	46 (30)	49 (34)	0-2	3-4	32
	TAIL-PINCH RESPONSE						
A	0	0-4	0-6	49 (60)	50 (57)	4-5	59%
B	2-16	75 (12)	0-18	0-34	0-2	2	12
C	0	0-10	81 (6)	16 (24)	0-3	3	15
D	0-12	37 (53)	43 (43)	18 (78)	0-6	2-3	58
E	0-2	32 (27)	15 (83)	49 (29)	0-10	2-4	46
F	0-13	74 (15)	0-3	19 (48)	0-5	2	32
G	0-18	67 (36)	0-48	24 (96)	0-3	2-4	66
H	0	0-25	0-2	85 (15)	0-34	4	15

[1]Data for 14 studies are represented for all laboratories except laboratory G (13 studies) and C (5 studies)
[2]Mean and C.V. of frequency distributions of the baseline control means, where the frequency >10.0%; otherwise range of frequencies are listed
[3]Range of baseline medians across studies
[4]Mean of C.V.s of the peaks of the baseline distribution where the frequency >10.0%

differed in replicability of these scores, and the grand C.V.s were mostly 32-56%, but ranged from 9% to 74%. Few '1's and '5's were ever observed in baseline data, or in control rats across time. In general, the distribution of scores remained about the same across testing, with only two laboratories showing slight decreases in reactivity.

Tail-Pinch Response. The tail-pinch response showed the least consistency of the sensorimotor endpoints, both between and within laboratories. The baseline control distributions, listed in Table 5 and presented in Figure 9, indicate considerable variability in these distributions between studies. Median response values in most studies were either '2' (in three laboratories), '3' (two laboratories), '4' (two laboratories) or '5' (one laboratory). In four laboratories, the percentage of rats receiving '2's and '4's was greater than those receiving a '3' (see Figure 9). Differences in baseline distributions were sometimes, but not always, obtained with different observers. In three laboratories, the grand C.V. was ≤15%, but replicability was 32-66% in the remaining laboratories. These discrepancies could be due to the variable nature of the stimulus, inconsistency in scoring the response, or inherent differences in groups of rats. As with the click response, '1's were rarely recorded, and only one laboratory (A) recorded a considerable proportion of '5's. Despite differences across studies, control values remained generally stable during the course of study.

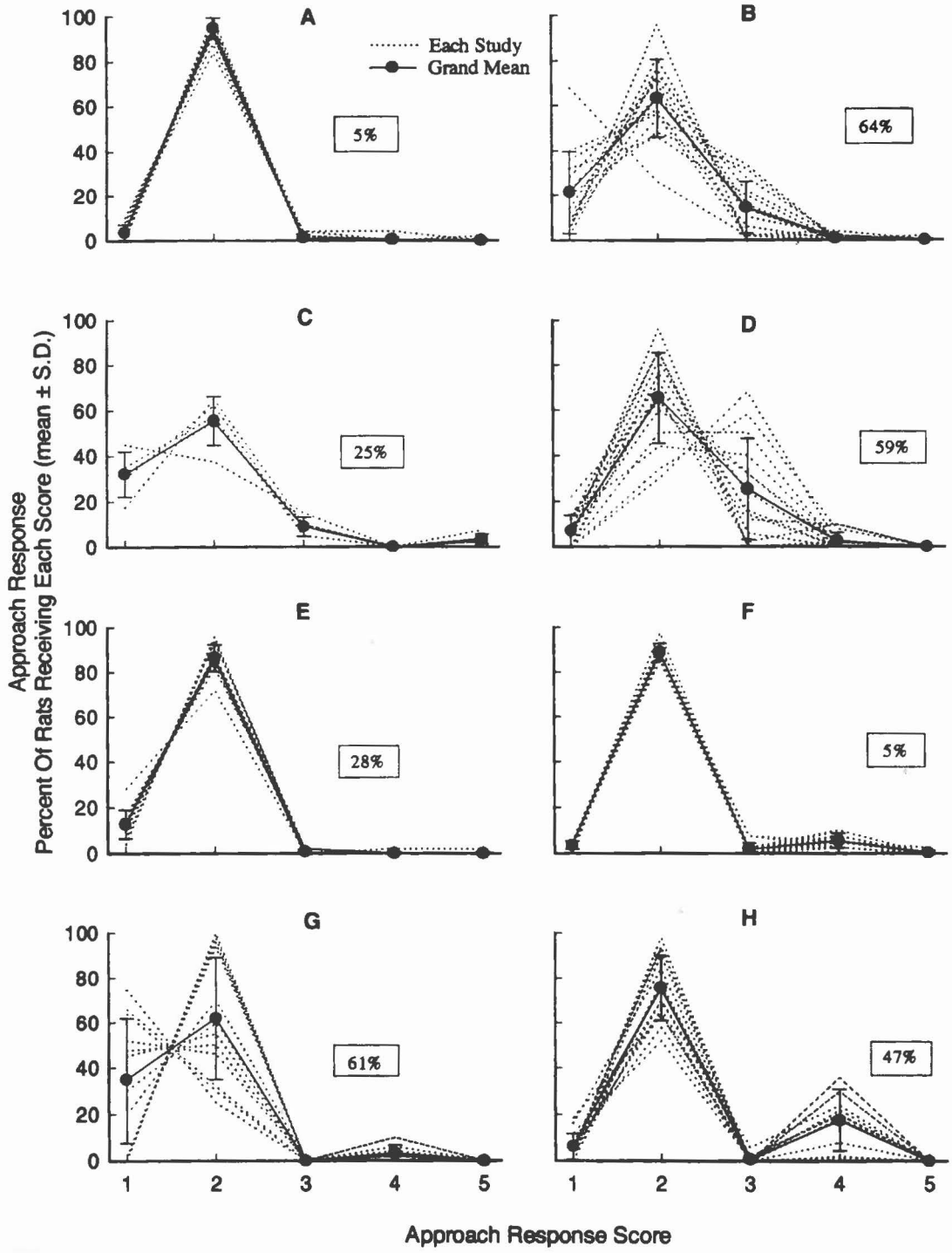

FIG. 8. Frequency polygons of the distributions of scores for the approach response, i.e., the percentage of rats receiving each of the possible scores (1 to 5). For each laboratory (A-H), the frequency polygon for the time-0 samples (n=40 or 50 rats/study) are plotted (dashed lines). The grand distributions, or mean ± standard deviation of the frequencies of each score across studies are also plotted (solid line). The grand C.V., i.e., mean of C.V.s of the major peaks of the grand distribution, are indicated for each laboratory.

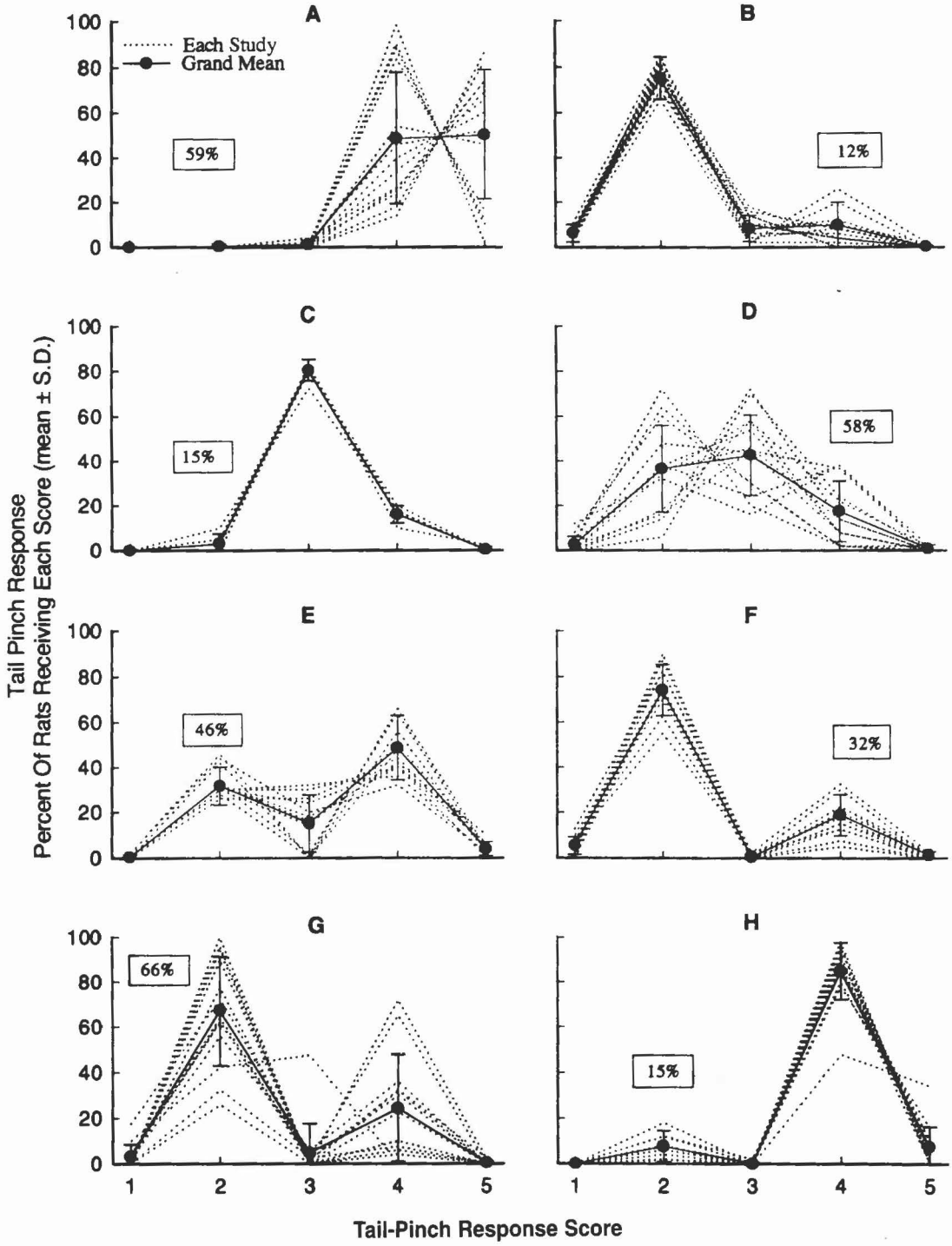

FIG. 9. Frequency polygons of the distributions of scores for the tail-pinch response, i.e., the percentage of rats receiving each of the possible scores (1 to 5). For each laboratory (A-H), the frequency polygon for the time-0 samples (n=40 or 50 rats/study) are plotted (dashed lines). The grand distributions, or mean ± standard deviation of the frequencies of each score across studies are also plotted (solid line). The grand C.V., i.e., mean of C.V.s of the major peaks of the grand distribution, are indicated for each laboratory.

TABLE 6. Baseline Control Data for Body Weight (grams) and Rectal Temperature (°C)[1].

Laboratory	Baseline means				C.V.s of baseline means[4]		
	Grand mean[2]	Grand C.V.	Range[3] low	Range[3] high	Mean	Range low	Range high
	BODY WEIGHT						
A	326	7%	290	372	5%	4%	6%
B	382	15	284	474	7	4	11
C	355	2	349	367	2	2	3
D	288	9	243	325	5	3	8
E	337	13	286	416	7	4	15
F	333	7	286	374	7	5	11
G	324	22	245	477	6	4	11
H	224	5	210	253	5	4	7
	BODY TEMPERATURE						
A	38.2	0.4%	37.8	38.3	1.0%	0.8%	1.2%
B	37.6	1.0	36.7	38.0	1.2	0.8	2.0
C	38.0	0.3	37.8	38.1	0.9	0.6	1.1
D	38.0	0.9	37.2	38.5	1.0	0.7	1.6
E	37.9	0.6	37.6	38.6	0.9	0.7	1.3
F	38.2	0.3	38.0	38.4	1.1	0.9	1.4
G	37.8	0.8	37.4	38.4	1.2	0.9	1.9
H	38.1	0.8	37.6	38.6	0.8	0.6	1.0

[1]Data for 14 studies are represented for all laboratories except laboratory G (13 studies) and C (5 studies).
[2]Mean of time-0 means for each laboratory, and C.V. of the baseline means
[3]Range of group means at time-0 for each laboratory
[4]Mean and range of C.V.s at time-0 for each laboratory

Other Measures

Body Weight and Temperature. Data for standard toxicity endpoints, body weight and temperature, are listed in Table 6. For body weight, some differences in mean values were evident across laboratories which were probably attributable to the different strains of rat; for example, laboratory H clearly used the smallest rats. Other differences could not be due to strain, however, since the laboratories using Sprague-Dawley rats had mean weights of 288-382 grams; this implies that the rats' ages may have differed at the start of the studies. Furthermore, some laboratories showed considerable differences in group means across studies (grand C.V.s up to 22%), also indicating that ages may have been somewhat different. Body temperature values were consistent within each laboratory, with average values between 37.6-38.2°C across laboratories and studies. Control variability of these physiological measures was generally low in all laboratories, with typical C.V.s from 2-7% for weight and 0.8-1.2% for temperature.

Piloerection. Piloerection was rarely recorded in control rats or at time-0 in five laboratories. Of the remaining three, laboratory A occasionally reported 10-30% of control rats showing piloerection at some time during a study. Laboratories B and C had numerous instances of piloerection at time-0, from 4-23% for specific studies, and an average of 11-13% of the rats across all studies. Furthermore, 10-50% of the control group was reported to show piloerection during the course of testing. In these laboratories, decreased piloerection in treated rats was sometimes obtained, but this was probably due to the high incidence in controls.

Reproducibility of the control data across laboratories is summarized in Table 7, which combines the overall means or medians of baseline data, overall replicability as presented by the inter-laboratory C.V.s, and overall control variability defined by C.V.s within each study. Clearly the least variable and most reproducible endpoint was body temperature, which hardly varied more than 1% of the group means. Body weight, landing foot splay, grip strength, and the subjective scoring of arousal, were also tightly con-

trolled within laboratories, although the actual control values (i.e., kilograms to release) varied more across laboratories. A different pattern was evident in the activity measures (motor activity and rearing), since in some laboratories the group data were highly variable (overall control variability of 36% and 44%) but the replicability of group means were more consistent (grand C.V.s of 15-16%). Scoring of the general reactivity (ease of removal, handling reactivity) and sensory-specific reactivity (approach, touch, click, and tail-pinch responses) endpoints were less reliable across studies, with grand C.V.s around 40%. While these high C.V.s may be due somewhat to the method used to derive replicability estimates, an assessment of Figures 5, 8, and 9 verify the high C.V.s. Control data for urinary and fecal output were highly variable and no estimates of these differences could be determined.

DISCUSSION

The sensitivity of these endpoints to the chemicals tested cannot be assessed on the basis of control data. Likewise, retest stability of control data is clearly not identical with the replicability of response to chemicals. Variability of control data, however, determines the statistical power of the test, and power of the test is one component of the sensitivity. Moreover, a low statistical power can sometimes be compensated for by increasing the numbers of animals. Test measures with a very low variability may sometimes be very insensitive to chemicals, and also high control variability may indicate high sensitivity of the measure to a variety of factors in addition to chemical treatment. This may be especially true for non-specific tests (e.g., the excitability measures). A very low variability of ordinal values may reflect more a stereotypical classification than stability of the criterion (e.g., arousal).

Several further important methodological problems could be addressed on the basis of the analysis of control data. Stability of data within the same control subject (retest correlation) would have important implications on two issues of study design: 1) standardization of the treatment groups according to baseline values; and 2) evaluation of main effects in relation to baseline values (e.g., covariance analysis). These approaches would be rational only for sufficiently stable variables. The effects of time of day and season of year, as well as order of testing, may have considerable influence on activity and excitability measures, and also on other variables which are modified by the subject's excitability (e.g., grip strength). In this Collaborative Study, no attempt was made to correlate control data variability or chemical effects with the order of testing, time of day, or season.

The neuromuscular endpoints were generally reproducible within most laboratories with low control variability, both at the beginning of and throughout studies. Indeed, some laboratories (especially laboratory C) showed such low control variability that even slight (e.g., 10%) differences between treatment groups and controls were statistically significant. Generally, these neuromuscular measures may be expected to be less variable due to the intense demands placed on these abilities for normal functioning. In addition, the motor output of the subject is integral to most of the tests of the neurobehavioral screening battery, and an assessment of this function is therefore critical to interpreting changes in the other endpoints.

The control variability of the rearing data, although high, was similar across laboratories and mean values were quite reliable across studies. In contrast, control variability for motor activity values differed greatly across laboratories. The C.V.s obtained in this study (20-53%) are somewhat higher than those reported in another comparison across laboratories (Crofton *et al.*, 1991), wherein the authors present C.V.s of historical control databases which range from 19% to 31% in five laboratories. Reasons for this discrepancy are unclear, but it does not appear to be due to the test apparatus, laboratory location or affiliation. High control variability decreases the chance of detecting statistically-significant changes in treated rats. This was indeed the case for laboratory G which never obtained significant motor activity changes (see Moser *et al.*, 1997b), and for which the variability of control groups after the initial test averaged 70% (ranging from 16-148%).

Both activity measures (rearing and motor activity) change with repeated testing, probably reflecting habituation and acclimation to the testing environment. The demonstration of greater decreases in the acute studies, in which test times were more closely spaced, supports this explanation. In contrast to the activity measures, the excitability measures (ease of removal, handling reactivity, arousal) did not clearly change with repeated testing. A decline in reactivity scores would be anticipated with ongoing dosing, testing and handling, yet the data from these laboratories demonstrate that this is not a predictable feature of these endpoints.

The removal and handling endpoints appear to measure approximately the same level of reactivity in control rats, in that laboratories generally showed the same ranking of excitability in both. Indeed, in almost all cases there was a good correlation between the ease of removal

TABLE 7. Overall Control Data Across all Laboratories, Presented In Order of Descending Replicability.

Continuous Data[1]	**Means**		**C.V.s Between Studies**		**C.V.s Within Studies**	
	Overall	Range	Overall	Range	Overall	Range
Body Temperature	38.0	37.6-38.2	0.6%	0.3-1%	1.0%	0.8-1.2%
Landing Foot Splay	78	52-92	8	2-14	20	14-28
Forelimb Grip	1.00	0.80-1.37	9	0-20	13	3-17
Body Weight	321	224-382	10	2-22	6	2-7
Hindlimb Grip	0.83	0.59-1.20	13	1-27	15	4-20
Rears	10	8-13	15	10-25	44	31-59
Motor Activity	-[3]	-	16	6-34	36	20-53

Ordinal Data[2]	**Medians**	**C.V.s between Studies**	
	Range	Overall	Range
Arousal	4	15%	2-24%
Approach Response	1-3	37	5-64
Tail-Pinch Response	2-5	38	12-66
Touch Response	2-3	40	8-73
Ease of Removal	1-3	41	11-70
Click Response	2-4	42	9-74
Handling Reactivity	1-4	42	8-67

[1]Grand means and C.V.s between and within studies

[2]Grand medians and C.V.s of frequency distributions

[3]Different units of measure were collected from different activity devices

and the handling reactivity scores; however, these did not correlate as well with the arousal measure. The arousal endpoint also showed the least variability across time points, studies, and laboratories. The observed variability across studies for ease of removal and handling reactivity could be due to a number of factors such as strain, age, and size of rat, but are probably most influenced by the amount and consistency of handling the rats before and during the study. Laboratory C, which recorded the lowest scores for these measures, stated that their experimental paradigm included extensive daily handling of the rats the week before testing began. Even within the same laboratory, personnel handle rats differently which may account for the between-observer variations evident in the data. There are also possible differences between rat shipments, and there were instances of particular studies in which all rats received atypical scores even with the same observer. These control data influence the ability of the endpoints to detect treatment-related changes. For example, laboratory C never detected decreased reactivity in these studies, most likely due to the very low control values. Significant differences between treatment and control groups may also be due to the fluctuations of the control data rather than to treatment; this may be suspected when there is no evident dose-response and the controls appear aberrant.

Most of the autonomic endpoints evaluate physiological functions (e.g., lacrimation) which are altered only by treatment or pronounced systemic toxicity. Excretory patterns, however, are influenced by many factors including the rat's reactivity (emotionality), physical state (food and water intake, diurnal cycle, metabolic changes), as well as autonomic variables (smooth muscle innervation). Counting urine pools and fecal boluses as indices of elimination was not as discrete as might be anticipated, and differences between observers were common. There were also differences between and within laboratories on the baseline control data and on the control data across time. This lack of stability and consistency probably accounts for the numerous instances in which significant differ-

ences were recorded between one or more treatment groups and control.

The approach and touch responses showed the lowest magnitude or strength of response across all laboratories, and also were generally more consistent across studies. For these measures it was not unusual for rats to receive a score of '1'. On the other hand, rats almost always showed some reaction to the click and tail-pinch stimuli, but these data (especially the tail pinch) were not as consistent from study to study. The relatively lower scores for approach and touch may account for the number of instances where increased reactivity was recorded for those measures, whereas the click and tail-pinch responses rarely showed increases. Indeed, for approach and touch responses, significant increases were obtained about as frequently as were decreases. Laboratories which recorded higher reactivity scores (especially the tail-pinch response in laboratory A) tended to have more instances of significantly-reduced reactivity in treated rats.

Thus baseline differences in control data influenced the results of the chemical testing (see Moser *et al.*, 1997b) for many measures in this Collaborative Study. In some cases the between-laboratory differences in these control data can be explained by information on the laboratories' general procedures (e.g., laboratory C extensively handled the rats before testing, and subsequently rated them as low reactivity on certain endpoints). Others, however, cannot be so easily explained. For example, grip strength values were considerably different across laboratories, even though all used essentially the same strain gauges. There were no obvious correlations between the grip strength values and the configurations or types of wire mesh or bars attached to the strain gauges. Likewise, there were differences in motor activity, even between laboratories using the same device. It should be noted that all participants in this Collaborative Study followed the same protocol, so it is possible that even more differences would be obtained when comparing data from laboratories using markedly divergent protocols. Finally, those endpoints which proved to be less reproducible across studies and laboratories are candidates for further modification.

ACKNOWLEDGEMENTS

The authors wish to thank Drs. J. Harry and M. Ehrich for their careful review of this manuscript. We also thank P. Phillips and J. Reese, who provided support for entering and checking these data. All data presented here are the result of considerable time, effort, and cost by project directors, laboratory personnel, and management in the participating laboratories, whose contributions must also be acknowledged.

REFERENCES

Crofton KM, Howard JL, Moser VC, Gill MW, Reiter LW, Tilson HA, MacPhail RC. Interlaboratory comparison of motor activity experiments: Implications for neurotoxicological assessments. *Neurotoxicol Teratol* 1991; 13:599-609

Moser VC, Tilson HA, MacPhail RC, Becking GC, Cuomo V, Frantík E, Kulig BM, Winneke G. The IPCS collaborative study on neurobehavioral screening methods: II. Protocol design and testing procedures. *Neurotoxicology* 1997a; 18:929-938

Moser VC, Becking GC, Cuomo V, Frantík E, Kulig BM, MacPhail RC, Tilson HA, Winneke G, Brightwell WS, De Salvia MA, Gill MW, Haggerty GC, Hornychová M, Lammers J, Larsen JJ, McDaniel KL, Nelson, BK, and Østergaard G. The IPCS collaborative study on neurobehavioral screening methods: V. Results of chemical testing. *Neurotoxicology* 1997b; 18:969-1056

MS #3225 *(Accepted: June 1, 1996)*

Volume 18, Number 4 / 1997

NeuroToxicology®
ISSN-0161-813X

INTOX PRESS, INC.
Publisher of NeuroToxicology
P. O. Box 24865
Little Rock, Arkansas 72221-4865
USA

(501) 227-8622 (Voice)
(501) 224-1947 (FAX)

PRINTED IN THE UNITED STATES OF AMERICA

NeuroToxicology® (ISSN-0161-813X) is published quarterly for $169 per year (institutional rate) and $79 per year (individual rate - if individual's institution subscribes) by Intox Press, Inc., 13924 Rivercrest Drive, Little Rock, Arkansas 72212. Additional $29 for shipping and handling outside the United States of America. Postmaster send address changes to *NeuroToxicology* c/o Intox Press Inc., P. O. Box 24865, Little Rock, AR 72221-4865 USA. Periodical Postage paid at Little Rock, Arkansas, and additional mailing offices.

This journal is printed on acid-free paper.

NeuroToxicology®

Volume 18 | 1997 | Number 4

Special Issue

THE IPCS COLLABORATIVE STUDY ON NEUROBEHAVIORAL SCREENING METHODS

Continued