Bull. Environ. Contam. Toxicol. (1992) 49:648–655

Environmental Contamination and Toxicology

Optimized Design for Earthworm Survival Tests in Soil

Stephen C. Sheppard and William G. Evenden

Environmental Science Branch, AECL Research, Whiteshell Laboratories, Pinawa, Manitoba R0E 1L0, Canada

Toxicity tests for soils have not been as extensively developed as those for aquatic systems (van Straalen and Denneman 1989); yet there is an urgent need to assess the biological impact of soils contaminated with unknown or mixed contaminants (van Leeuwen 1990). A dilemma is caused by the profound effect that soil properties, quite apart from the presence of a contaminant, can have on test organisms. For this reason, many protocols call for the use of liquid soil extracts rather than the whole soil. Unfortunately, organisms also respond to particle-bound contaminants and can play a role in the release of these contaminants. Therefore, there is a need to assess toxicity in whole soils. Assays with earthworms have been developed, although many of them involve artificially contaminated soils (e.g., Pizl 1988) and often artificial soils as well (e.g., Neuhauser et al. 1986). Some modification is required to use these assays for whole soils from contaminated sites where the contaminants may or may not be known.

Several researchers have sought to use the growth and reproduction of earthworms in toxicity tests (Lofs-Holmin 1980; van Gestel et al. 1988, 1989). Certain species are well suited to laboratory rearing. Others have longer life cycles, so that measures of growth and reproduction are time consuming and often impractical (Butt 1991; Iglisch and Kriegerowski 1986; Lofs-Holmin 1980). *Lumbricus* species are among these; they would otherwise be well suited to toxicity assays because of their broad ecological and geographical distributions and importance as decomposers. Large numbers of *L. terrestris* are available as commercial angling bait; 10^9 worms worth $50 million are picked annually in Ontario, Canada (Tomlin and Protz 1990). Although these sources may not provide strictly the one species, the uniformity and continuity of supply are distinct advantages, and these worms have been used for toxicity assessment in Germany (Heimbach 1985) and on Superfund sites in the U.S. (Callahan et al. 1991).

The objective of this paper is to present the results of a study on the optimal experimental designs for survival tests using bait worms. We found no reports in the literature addressing this subject for earthworms. The most common experimental design with earthworms uses 1 to 10 worms per container and 3 to 5 replicates (e.g., Pizl 1988; Heimbach 1985; Neuhauser et al. 1986). Experience in our laboratory indicated this arrangement was not satisfactory (Sheppard et al. 1992), and this study was initiated to investigate the alternatives.

Send reprint request to S.C. Sheppard.

MATERIALS AND METHODS

The objective of Experiment A was to determine the number of replicates per soil needed to indicate significant differences in survival across a series of treatments, in this case soil types. Eleven soils were chosen to represent a broad range of soil properties (Table 1). The soils were partially dried to a workable moisture content, sieved to pass a 0.5-cm mesh, and placed in 500-mL plastic containers with soil surface areas of 90 cm^2. Sixteen containers of each soil were used. The soils were moistened to field capacity, and five mature worms (*L. terrestris*), selected for uniformity of size, were added. The worms had been freshly obtained from a commercial angling-bait supplier in Ontario and were relatively uniform in size and apparent maturity. After 54 d at 15°C, the remaining live worms were extracted by hand and counted. To assess the optimal number of replicates, the observations from Experiment A were randomly sorted six times. Each time an analysis of variance (ANOVA) was calculated using subsets with 3, 5, 8, 10 or 16 observations for each soil.

The objective of Experiment B was to compare means of improving the precision of survival tests, with the underlying hypothesis being that increased numbers of earthworms or experimental units would be useful. Preliminary trials indicated that placing earthworms with soil in plastic bags was expedient, and that packaging worms singly in this manner was practical. The bags were made of polyethylene tubing, 9-cm or 25-cm diameter when flat, heat-sealed to form small bags. The size of the bag allowed the soil to be spread to a 2-cm-thick layer. Each bag was punctured with about ten holes, using a hypodermic needle, to ensure sufficient air exchange.

Table 1. Properties of the soils used in Experiment A, ordered by clay content, and showing mean (± standard deviation) bait worm survival (5 worms per container, n = 16) and fraction of units with all worms dead.

Soil[a]	Texture[b]	Clay (%)	pH[c]	Organic matter content (%)	Survival[d] (%)	Fraction of units with all dead (%)
A1	CL	33	6.3	3.1	88 ± 34	13
A2	L	24	7.4	2.2	98 ± 5	0
G	fine SL	18	7.1	18.4	50 ± 47 *	44
A3	fine SL	15	7.5	2.6	56 ± 51 *	44
A4	fine SL	13	6.4	5.7	100 ± 7	0
A5	fine SL	12	7.8	4.2	82 ± 38	13
B1	LS	6	5.8	3.5	74 ± 44	25
A6	L fine S	4	6.2	0.8	68 ± 48	31
IB2	fine S	2	5.9	1.0	24 ± 43 *	75
uB2	fine S	2	5.1	0.7	26 ± 42 *	69
O	O	<1	5.4	41.5	34 ± 46 *	60

[a] The symbols refer to agricultural soils (A1-A6), humus enriched garden soil (G), boreal forest podzolic sands (B1, IB2-limed, and uB2-unlimed) and organic (O).

[b] Symbols denoting texture are C (clay), L (loam), O (organic) and S (sand).

[c] pH measurements shown were taken after the experiment.

[d] The overall F ratio from analysis of variance for the effect of soil was significant at $P \leq 0.0001$ when all 16 replicates were included, and means indicated by * were significantly below 100%.

The treatments (Table 2) included a series at a constant density of 80 cm^3 of soil per worm, with 1, 5, 10 and 50 worms per container. Replication of these units was 100, 20, 10 and 2, respectively. The 4000 cm^3/50 worm treatment was held in plastic pails, rather than bags, to reduce handling injuries to the worms. Another series of a decreased soil-to-worm ratio using the same volumes of soil and numbers of replicates, ranging from 40- to 20-cm^3 of soil per worm. A commercial potting soil moistened to field capacity was used. Earthworms were obtained and selected for use as before. For the treatments that involved one or two worms per container, the worms were weighed as they were placed into the soil, and again after they were removed. After the worms were added to the soil, the containers were incubated at 15°C for 28-30 d. Moist absorbent paper was placed over the bags to reduce moisture loss through the aeration holes. After 14 d, counts of live worms in the bags were done by observing them through the plastic. At the end of the incubation, live worms in all containers were counted.

An ANOVA was done on the full Experiment B to determine the significance of the treatments on survival. The data for each treatment had different expected frequency distributions. The data were transformed to compute non-parametric tests by calculating the surviving fraction for each treatment, then assigning median scores, where survival fractions above the experiment median were assigned the score 1, and fractions below the median were assigned the score 0. The scores were used to compute the ANOVA, and treatments were compared by single degree of freedom contrasts.

To investigate the optimal number of replicates, different numbers of observations for each treatment in Experiment B were randomly selected, the means calculated, and the results compared to the overall treatment means. This was repeated in six independent cases for each level of replication, and the optimal number of replicates was estimated as the number where all six cases had means within five percentage points of the overall treatment mean.

Simulated data, based on the frequency distributions observed in Experiment B, were used to compare the statistical power of various experimental design options. The options considered were experimental units that contained one worm, where the data are binary survival scores, and experimental units that contained ten worms, where the data may approximate a normal or bimodal frequency distribution. It was anticipated that these approaches would require markedly different numbers of units and worms to achieve the same level of statistical precision. The issue of parametric versus non-parametric statistical methods also arises. We used non-parametric methods based on median scores, unless otherwise specified.

Simulated control data were compared with simulated treatment data that had half the number of survivors of the controls, equivalent to the concentration lethal to half the sample (LC_{50}) frequently sought in toxicity trials. Two control survival regimes were used. One used the survival rate observed in Experiment B with one-worm units (88%). This was simulated by randomly assigning values of 0 or 1 to a set of 1000 observations, with the target of 120 observations set to 0 (dead) and 880 set to 1 (live). The corresponding simulated LC_{50} observations had a target of 44% survival. The other control survival regime simulated the mean survival and survival count frequency distribution observed in Experiment B with ten-worm units. Here the mean survival was 75%, with 20% of the units with no survivors, 10% with 5 survivors, and 70% with 10 survivors. The corresponding simulated LC_{50} observations had a target of 37.5% survival. The frequency distribution of the simulated LC_{50} observations was designed to mimic that of the controls, and therefore was bimodal.

Table 2. Worm survival at 14 and 30 d in Experiment B, which evaluated the most effective means to increase the number of worms used in an assay.

Treatment	n	Soil per worm (cm^3)	Worms per unit	Type of container	Survival at 14 d meanscore[a]	Survival at 14 d (%)	Survival at 30 d meanscore	Survival at 30 d (%)	Fraction of units with many dead[c] (%)	To obtain consistent results[d] Number of units	To obtain consistent results[d] Number of worms
High ratio	1.	100	80	1	bag	88	0.88[b]	88	0.88	12	56	56
	2.	20	80	5	bag	90	0.80	87	0.80	10	16	80
	3.	10	80	10	bag + tray	75	0.70	75	0.70	20	8	90
	4.	2	80	50	bucket	---	---	75	0.50	---	---	---
Low ratio	5.	100	40	2	bag	93	0.92	92	0.92	8	50	100
	6.	20	33	12	bag	88	0.70 *[b]	85	0.70 *	15	15	180
	7.	10	27	30	bag + tray	68	0.60 *	59	0.50 *	40	9	280
	8.	1	20	200	bucket	---	---	10	---	---	---	---

[a] Means are arithmetic mean survival fractions, scores are the median scores used in non-parametric tests of significance where the scores were 1 for survival fractions above the treatment median, and 0 for fractions below the treatment median.

[b] Tests of significance contrasted the score of each treatment versus the score of treatment #1, with $P < 0.05$. The score for treatment 8 was meaningless because $n = 1$.

[c] The fraction of units with many worms (>80%) dead indicates the incidence of cascade deaths.

[d] The number of units and worms to obtain consistent results was determined by randomly selecting different-sized subsets of data from each treatment. Results were considered consistent when the survival in the subsets was no more than five percentage points different from the survival mean of all replicates.

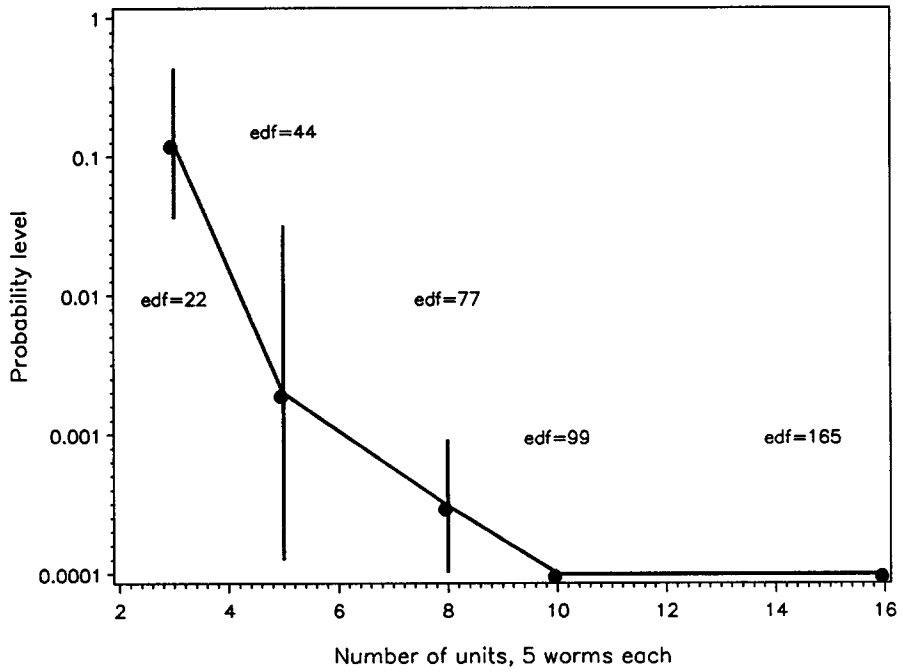

Figure 1. Probability (P) for an effect on earthworm survival of soil type, among 11 soils in Experiment A, as influenced by the number of replicates considered. Error bars are ± one standard deviation for six random selections of replicates from the full 16 replicates of the experiment. P was based on one-way ANOVAs, and the corresponding number of error degrees of freedom (edf) are indicated.

Each simulation case was reproduced five times, each with independent, randomized generation of observations. Conclusions were based on the calculated probability levels for a difference between the control and LC_{50} treatments, based on ANOVA. The analysis of binary data by ANOVA is equivalent to non-parametric analysis of median scores. The bimodal data from the ten-worm units was analysed both as parametric and as non-parametric using median scores.

RESULTS AND DISCUSSION

Survival was significantly different among the soils in Experiment A (Table 1), with the poorest survival in acidic sandy soils. Neutral sands and slightly acidic heavier textured soils did not markedly reduce survival. Soils that had poor survival would be difficult to use in toxicity tests, because the impact of the contaminant is assessed on the numbers of live worms in uncontaminated controls.

Perhaps the most important observation in this experiment is the occurrence of cascade deaths. The data were clearly bimodal, with most containers having either all dead or all living worms. This is reflected in the fractions of units with all dead (Table 1), which almost fully account for the survival data. Evidently, when worms are stressed by soil conditions, soil-borne disease or toxicants, it is likely that all the worms in the container will suffer and, if one worm dies, it is likely to precipitate the deaths of others. For this reason, individual packaging has apparent advantages.

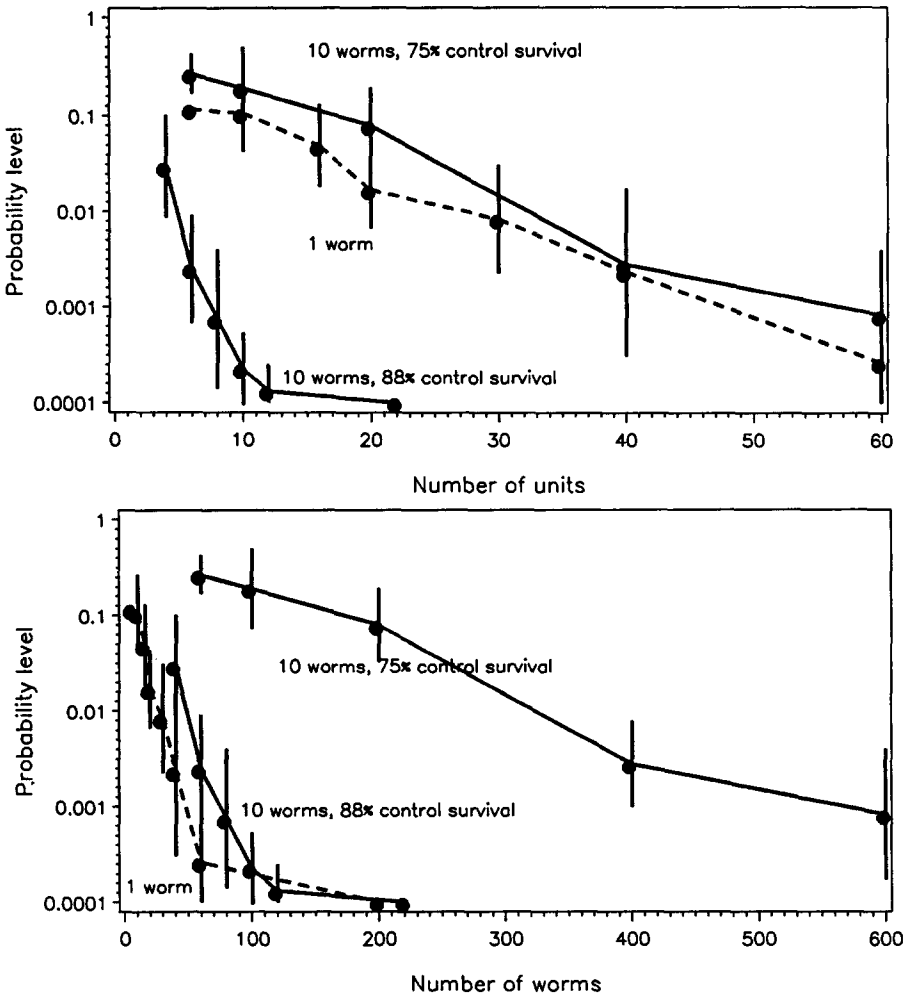

Figure 2. Probability (P) for a significant difference between controls and a treatment where half the earthworms died (equivalent to an LC_{50} treatment), plotted versus the number of experimental units (a) and versus the number of worms (b). Shown are the results for the one-worm units, which had 88% control survival, ten-worm units with an 88% control survival, and ten-worm units with a 75% control survival. Error bars are ± one standard deviation for five sets of simulated data, as described in the text.

There was a four-fold range in survival among soils in Experiment A (Table 1). The number of replicates required to consistently show significant differences among the soils was 5 to 8 (Figure 1). About 10 replicates gave as much consistency among the means and as much statistical precision in ANOVA as the full 16 replicates. These ANOVA involved 11 soils, so that there were 44 error degrees of freedom (edf, Figure 1) when 5 replicates were considered. For a test between only two treatments, it would require 23 replicates to obtain the same number of error degrees of freedom. These numbers of replicates and implied numbers of worms are well above those commonly used, and may present a logistic constraint in laboratory use. Because of this, we sought more efficient experimental designs.

There were marked effects of grouping worms and decreasing the soil-to-worm ratio in Experiment B (Table 2). Survival was optimal in containers that held one or two worms. The proportion of cascade deaths increased as more worms were used per container. Evidently, using multiple worms per container decreased survival enough to jeopardize the estimation of effects such as LC_{50}. These data strongly support the use of containers with only one or two worms. However, the implications for statistical power are not clear, and this led to our simulation investigations.

We also used the data from Experiment B to determine the optimal number of replicates to obtain consistent results for each treatment (Table 2). For the one- and two-wormtreatments, about 50 to 56 replicates gave consistency comparable to 100 replicates. After 30 replicates, the improvements in consistency were small. In contrast, about 9 replicates were optimal for treatments with 10 to 30 worms per unit, implying the need for 90 to 270 worms.

The worms lost about 13% of their original live weight in the one- and two-worm treatments of this experiment. In a preliminary experiment, also using one worm per bag, they gained about 20%. The weight gains and losses within each trial were fairly consistent, with standard deviations of about 7% and 16% of the means. However, a method is needed to allow correction for the weight of soil in the gut. We did preliminary investigations contacting worms with varying concentrations of phenolphthalein (a common purgative for mammals),ethanol and methanol, but were unable to accelerate depuration of soil.

Using the simulated data, we investigated the number of replicates required to be able to assign statistical significance to the simulated LC_{50} treatment. In simulations with an 88% control survival, LC_{50} was significant at the 0.01 probability level when there were 5 replicates with 10 worms per unit, or when there were 27 replicates with 1 worm per unit (Figure 2). The former is efficient to minimize numbers of units, the latter to minimize numbers of worms. However, our experimental results indicate that control survival decreases when 10 worms are grouped together (Table 2), and when the observed control survival for 10 worms per unit was simulated, the number of replicates to obtain the 0.01 probability level rose to 29 (33 if analysed as parametric data). This is inefficient both for numbers of units and numbers of worms when compared to the one worm per unit case. We conclude that to assign significance to LC_{50} effects, the one worm per unit design is superior.

Survival tests in soil using *L. terrestris* have been reported in the literature, but many of the reports involve few replicates, and we found no reports that investigated the optimal experimental design. Optimal designs should maximize the statistical power of comparisons and minimize the numbers of worms and experimental units required, and the costs. Our first experiment involved experimental units with five worms each, and we observed that 5 to 10 replicates were required to give reasonable precision. It was also apparent that cascade deaths were a problem, so that most units had either all alive or all dead. An obvious solution to the problem of cascade deaths is to use units with one worm each. The method we developed was convenient and low-cost, it allowed us to weigh individuals before and after exposure to the soil, and it allowed survival scoring at various times without disturbance. To obtain good statistical precision, it required 30 to 50 replicates. In contrast, experiments with units that contain 5 to 10 worms needed somewhat fewer replicates, but markedly more worms. Through our experiments and using simulated data to compare statistical power, we conclude that the optimal design involves large numbers of replicates of units with one worm each.

Acknowledgments. Funding for most of this project was provided by the Ontario Ministry of the Environment. The technical efforts of Jeanette Evans are appreciated. Helpful comments on drafts of the manuscript were provided by Glen Bird, Marius Marsh, Marsha Sheppard and Reto Zach.

REFERENCES

Butt KR (1991) The effects of temperature on the intensive production of *Lumbricus terrestris* (Oligochaeta: Lumbricidae) Pedobiologia **35**:257-264

Callahan CA, Menzie CA, Burmaster DE, Wilborn DC, Ernst T (1991) On-site methods for assessing chemical impact on the soil environment using earthworms: A case study at the Baird and McGuire Superfund site, Holbrook, Massachusetts. Environ Toxicol Chem **10**:817-826

Heimbach F (1985) Comparison of laboratory methods, using *Eisenia foetida* and *Lumbricus terrestris*, for the assessment of the hazard of chemicals to earthworms. Z Pflanzenkr Pflanzenschutz **92**:186-193

Iglisch VI, Kriegerowski L (1986) Indications for mass breeding of *Lumbricus rubellus* Hofm. (Annelida, Lumbricidae) as a representative test organism of terrestrial biotopes for the evaluation of chemicals. Anz. Schädlingskde. Pflansenschutz Umweltschutz **59**:95-99

Lofs-Holmin A (1980) Measuring growth of earthworms as a method of testing sublethal toxicity of pesticides. Swedish J Agric Res **10**:25-33

Neuhauser EF, Loehr RC, Malecki MR (1986) Contact and artificial soil tests using earthworms to evaluate the impact of wastes in soil. In: Petros JK, Lacy WJ, Conway RA (eds.) Hazardous and Industrial Solid Waste Testing: Fourth Symposium. American Soc Testing Materials ASTM STP 886, Philadelphia, p192-203

Pizl V (1988) Interactions between earthworms and herbicides. I. Toxicity of some herbicides to earthworms in laboratory tests. Pedobiologia **32**:227-232

Sheppard SC, Evenden WG, Anderson AJ (1992) Multiple assays of uranium toxicity in soil. Environ Toxicol Water Qual **7**:(in press)

Tomlin AD, Protz R (1990) Soil faunal activity and soil microfabrics in sustainable cropping systems. Proc Entomol Soc Ontario **121**:25-41

van Gestel CAM, van Dis WA, van Breemen EM, Sparenburg PM (1988) Comparison of two methods for determining the viability of cocoons produced in earthworm toxicity experiments. Pedobiologia **32**:367-371

van Gestel CAM, van Dis WA, van Breemen EM, Sparenburg PM (1989) Development of a standardized reproduction toxicity test with the earthworm species *Eisenia foetida andrei* using copper, pentachlorophenol, and 2,4-dichloroaniline. Ecotoxicol Environ Safety **18**:305-312

van Leeuwen K (1990) Ecotoxicological effects assessment in the Netherlands: Recent developments. Environ Manag **14**:779-792

van Straalen NM, Denneman CA (1989) Ecotoxicological evaluation of soil quality criteria. Ecotoxicol Environ Safety **18**:241-245

Received March 17, 1992; accepted May 5, 1992.