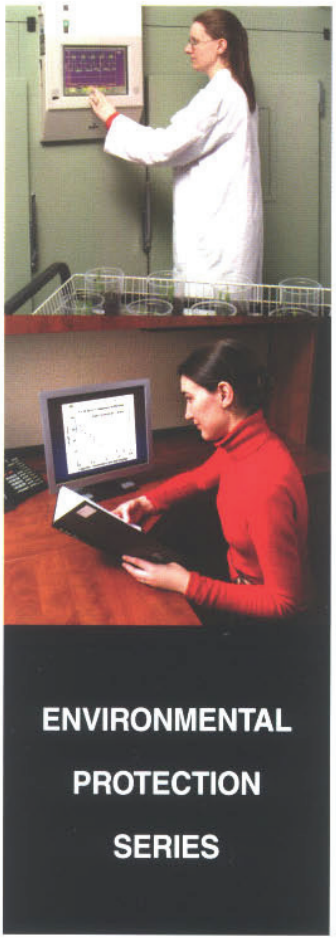
EPS 1/RM/46 – March 2005 **(with June 2007 amendments)**
Method Development and Applications Section
Environmental Technology Centre
Environment Canada



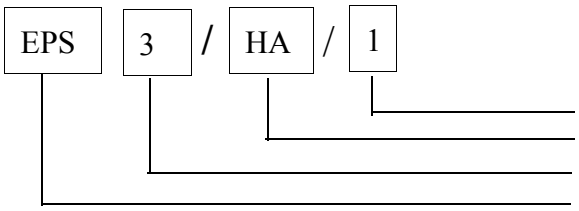
Guidance Document on Statistical Methods for Environmental Toxicity Tests

Environment Canada Environnement Canada

Canada

Environmental Protection Series

Sample Number:



Categories

1 Regulations/Guidelines/Codes of Practice
2 Problem Assessments and Control Options
3 Research and Technology Development
4 Literature Reviews
5 Surveys
6 Social, Economic and Environmental Impact Assessments
7 Surveillance
8 Policy Proposals and Statements
9 Manuals

Subject Areas

AG Agriculture
AN Anaerobic Technology
AP Airborne Pollutants
AT Aquatic Toxicity
CC Commercial Chemicals
CE Consumers and the Environment
CI Chemical Industries
FA Federal Activities
FP Food Processing
HA Hazardous Wastes
IC Inorganic Chemicals
MA Marine Pollutants
MM Mining and Ore Processing
NR Northern and Rural Regions
PF Paper and Fibres
PG Power Generation
PN Petroleum and Natural Gas
RA Refrigeration and Air Conditioning
RM Reference Methods
SF Surface Finishing
SP Oil and Chemical Spills
SRM Standard Reference Methods
TS Transportation Systems
TX Textiles
UP Urban Pollution
WP Wood Protection/Preservation

New subject areas and codes are introduced as they become necessary. A list of EPS reports may be obtained from Environmental Protection Publications, Environment Canada, Ottawa, Ontario, K1A 0H3, Canada.

Guidance Document on Statistical Methods for Environmental Toxicity Tests

Method Development and Applications Section
Environmental Technology Centre
Environment Canada
Ottawa, Ontario

Report EPS 1/RM/46
March 2005 (with June 2007 amendments)

Library and Archives Canada Cataloguing in Publication

Guidance document on statistical methods for environmental toxicity tests/
Method Development and Applications Section, Environmental Technology Centre,
Environment Canada.

(Report ; EPS 1/RM/46)
Includes abstract in French.
Issued also in French under title: Document d'orientation sur les méthodes
statistiques applicables aux essais d'écotoxicité.
Available also on the Internet.
One of a series of supporting guidance documents published by Environment
Canada that relate to recommended or standardized biological test methods. Cf. Foreword.
Includes bibliographical references: p. 280
ISBN 0-660-19509-7
Cat. no.: En49-7/1-46E

1. Toxicity testing--Statistical methods.
2. Environmental toxicology--Statistical methods.
3. Toxicology, Experimental--Statistical methods.
4. Biological assay--Statistical methods.
5. Water quality bioassay--Statistical methods.
6. Toxicity testing--Canada--Statistical methods.
I. Canada. Environment Canada
II. Environmental Technology Centre (Canada). Method Development and
Applications Section
III. Title: Biological test method.
IV. Series: Report (Canada. Environment Canada) EPS 1/RM/46.

QK46.5 S7 B56 2005 615.9'02'0727 C2005-980197-2

Catalogue No. EN49-71-46E
ISBN 0-660-19509-7

Readers' Comments

Comments regarding the content of this report should be addressed to:

Richard P. Scroggins
Chief, Biological Methods Division
Environmental Technology Centre
Environment Canada
335 River Road
Ottawa, Ontario
K1A 0H3

Cette publication est aussi disponible en français. Pour l'obtenir, s'adresser à:

Publications de la Protection de l'environnement
Environnement Canada
Ottawa (Ontario)
K1A 0H3

Review Notice

This report has been reviewed by the staff of the Environmental Technology Advancement Directorate, Environment Canada, and approved for publication. Mention of trade names or commercial products does not constitute endorsement by Environment Canada for use. Other products of similar value are available.

Abstract

This guidance document supports and supplements the methods for single-species toxicity tests, published by Environment Canada. In particular, it is intended for new laboratory personnel.

This document provides additional guidance for statistical analysis of results from Environment Canada tests. It comments on desirable procedures and common pitfalls. Some statistical background is covered, but this document does not teach basic statistics. Nor does it attempt to break new ground in statistical analysis, although it points to methods that are under development and seem promising for future use. This document covers methods for lethal and sublethal tests, with most emphasis on the more numerous aquatic tests (water-column and sediment).

A detailed glossary is provided. A design chapter emphasizes the need for consultation with a statistician, choice of concentrations, staying with logarithm of dose, the various types of controls, reference toxicants, randomization, replication, and transformation of data.

Choices among single-concentration tests are outlined, and the limitations imposed by design.

A section on quantal tests outlines methods for estimating effective concentrations (ECp) and confidence limits, and dealing with control effects. Various analytical methods provide similar endpoints for good data. Probit regression is recommended if there are two partial effects, preferably by maximum likelihood techniques. The Spearman-Kärber method with limited trimming is the choice if there is only one partial effect, and the binomial method if only zero and complete effects are available. A line should be plotted by hand to check for errors. Toxicity curves and analyses of effective times are beneficial.

For quantitative tests, which are usually sublethal, a point-estimate of the inhibition concentration (ICp) by regression is the most favoured method. Environment Canada has recently required linear and nonlinear regression as the first choice for analysis (Section 6.5.8). That analysis replaces the estimation of ICp by smoothing and interpolation (the ICPIN program) which has been commonly used. Hypothesis testing to determine a "no-observed-effect" concentration (NOEC) is outlined in detail because it has been used so frequently; this approach is much less desirable than point-estimates, and its use is decreasing.

In dual-effect tests, the correlation between the two effects, and their different statistical distributions, creates severe analytical problems. The most expedient approach is to separate the analysis of the quantitative component (usually sublethal) from the analysis of the quantal effect (usually lethal). An alternative approach that can sometimes be justified on ecological grounds is to combine the two effects into a "biomass" analysis, an approach that usually produces a more pronounced effect.

The statistical background includes discussion of difficulties caused by the customary "inverse" estimation of endpoints and confidence limits. Limited methods are described for testing significant differences between and among endpoints, and dealing with outliers. Advice is given for interpreting other deviant dose-effect relationships.

Résumé

Le présent document d'orientation étaye et complète les méthodes d'essai toxicologique monospécifique publiées par Environnement Canada. Il s'adresse en particulier au nouveau personnel de laboratoire.

Le document fournit des indications supplémentaires sur les analyses statistiques des résultats des essais mis au point par Environnement Canada. Il renferme des observations sur les procédures souhaitables et les pièges courants. Il présente certaines notions statistiques, mais ne comporte pas d'éléments de formation en statistique de base. Il ne tente pas non plus d'innover dans le domaine de l'analyse statistique, bien qu'il fasse état de certaines méthodes qui sont en cours d'élaboration et qui semblent prometteuses. Le document décrit les méthodes applicables aux essais de toxicité létale et sublétale, l'accent étant surtout mis sur les essais aquatiques (colonne d'eau et sédiment), qui sont plus nombreux.

Outre un glossaire détaillé, le document renferme une section sur la conception des essais. Cette section souligne l'importance que revêtent la consultation d'un statisticien, le choix des concentrations, le respect du logarithme de la concentration, les divers types de témoins, les toxiques de référence, la randomisation, les répétitions, la transformation des données.

Les essais à concentration unique parmi lesquels on peut choisir sont présentés, de même que les limites associées à la conception de ces essais.

Une section sur les essais visant à mesurer les effets quantiques décrit les méthodes d'estimation des concentrations efficaces et des limites de confiance, de même que la façon de tenir compte des effets observés chez les organismes témoins. Diverses méthodes d'analyse fournissent des résultats semblables et permettent d'obtenir de bonnes données. En présence de deux effets partiels, il est recommandé d'utiliser la méthode de régression des probits, de préférence à l'aide de techniques du maximum de vraisemblance. La méthode de Spearman-Kärber avec équeutage limité est à privilégier si l'on obtient un seul effet partiel, tandis que la méthode binomiale sera utilisée si seuls des effets zéro et extrêmes sont obtenus. Il convient de tracer à la main une courbe point par point afin de détecter toute erreur. Les courbes de toxicité et les analyses des temps efficaces présentent des avantages.

Pour les essais visant à mesurer les effets quantitatifs, qui sont habituellement des essais de toxicité sublétale, une estimation ponctuelle de la concentration inhibitrice (CI_p) par régression constitue la méthode la plus recommandée. Environnement Canada exige depuis peu de temps que la régression linéaire et non linéaire soit privilégiée pour les analyses (v. 6.5.8). Cette analyse remplace l'estimation de la CI_p par lissage et interpolation (programme ICPIN) utilisée couramment jusqu'à maintenant. La vérification d'hypothèse pour déterminer la concentration « sans effet observé » est décrite en détail du fait qu'elle a été employée très souvent; cette approche est nettement moins souhaitable que l'estimation ponctuelle et son utilisation diminue.

Dans les essais visant à mesurer deux effets, la corrélation entre ces derniers et leurs répartitions statistiques respectives occasionnent de graves problèmes analytiques. L'approche la plus pratique consiste à séparer l'analyse de l'élément quantitatif (effet habituellement sublétal) de l'analyse de l'élément quantique (effet habituellement létal). Une autre approche, qui peut parfois être justifiée d'un point de vue écologique, consiste à combiner les deux effets dans une analyse de la « biomasse »; cette approche se traduit généralement par un effet plus marqué.

Les notions statistiques incluent une analyse des difficultés que soulève l'habituelle estimation « inversée » des résultats et des limites de confiance. Le présent document décrit des méthodes restreintes permettant de vérifier les écarts significatifs entre les résultats et à l'intérieur de ceux-ci et de traiter les aberrations. Il fournit également des conseils sur l'interprétation d'autres relations dose-effet aberrantes.

Foreword

*This is one of a series of supporting **guidance documents** published by Environment Canada (EC), that relate to recommended or standardized biological test methods. The tests use single species of aquatic or terrestrial organisms under defined and controlled laboratory conditions, to measure adverse toxic effects from samples of selected materials. The recommended methods have been evaluated by Environment Canada and are favoured:*

- *for use in EC laboratories for environmental toxicity;*
- *for testing which is contracted out by Environment Canada, or requested from outside agencies or industry; or*
- *as a foundation for providing very explicit instructions which might be required in a standardized regulatory or reference method.*

The different types of tests in the series were selected to be suitable for the needs of environmental protection and management programs carried out by Environment Canada. The reports describing the test methods are intended to guide and facilitate the use of consistent, appropriate, and comprehensive procedures for obtaining data on toxicity to aquatic and terrestrial organisms. The tests are intended to be suitable for assessing simple or complex materials that are destined for release into the environment, or are already in some component of the environment such as sediment.

*Appendix A lists the generic (universal) multi-purpose **biological test methods**, the standardized **reference methods**, and the supporting **guidance documents**, which have been published to date. These reports, produced by Environment Canada's Method Development and Applications Section in Ottawa, Ont., are available from Environmental Protection Publications, Environment Canada, Ottawa, Ont., K1A 0H3, Canada. The guidance in the documents is shared and applied by the Regional and Headquarters Offices of Environment Canada (see Appendix C for contact information).*

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

APHA	American Public Health Association
ASTM	American Society for Testing and Materials
α	alpha (Greek)
ß	beta (Greek)
CAEAL	Canadian Association for Environmental Analytical Laboratories Inc.
CCREM	Canadian Council of Resource and Environment Ministers
CV	coefficient of variation
d	day(s)
EC	Environment Canada
EC50, ECp	median effective concentration, effective concentration for p% of individuals
EPS	Environmental Protection Service
g	gram(s)
GLIM	generalized linear model
GLM	general linear model
GLP	Good Laboratory Practice
h	hour(s)
IC, ICp	inhibiting concentration, inhibiting concentration for a reduction in performance of p% compared to the control performance.
ISO	International Organization for Standardization
kg	kilogram(s)
L	litre(s)
LC50	median lethal concentration
LOEC	lowest-observed-effect concentration
LSD	Least Significant Difference test of R.A. Fisher
mg	milligram(s)
MLE	maximum likelihood estimation (or estimate)
MSD	minimum significant difference
NOEC	no-observed-effect concentration
OECD	Organisation for Economic Co-operation and Development
‰	parts per thousand
QA/QC	Quality Assurance/Quality Control
R^2	coefficient of determination
s	second(s)
SD	standard deviation
SE	standard error
S-K	Spearman-Kärber method of analysis for quantal tests
∑	sigma (Greek)
TOEC	threshold-observed-effect concentration
USEPA	United States Environmental Protection Agency
s^2	variance
χ, χ^2	chi, chi-square (Greek)

Glossary

All definitions are given in the context of the procedures in this report, and might not be appropriate in another context. Words in italics, within a definition, are defined separately.

Grammatical Terms

Must is used to express an absolute requirement.

Should is used to state that the specified condition or procedure is recommended and ought to be met if possible.

May is used to mean "is (are) allowed to."

Can is used to mean "is (are) able to."

Might expresses a possibility that something could exist or happen.

Technical Terms

Accuracy is the closeness of the measured (or estimated) value to the "true" value. In toxicity testing, there can be no measure of accuracy because there is no way of knowing the "true" value of toxicity. (See *precision*.)

Acute means within a short period (seconds, minutes, hours, or a few days) in relation to the life span of the test organism.

Acute toxicity is a discernable adverse effect (lethal or sublethal) induced in the test organisms within a short period of exposure to a test material, usually a few days for larger organisms.

Algorithm signifies a set of rules for solving a problem. Historically, the term referred in a general way to arithmetic systems. Today, it is mostly used in the context of solving mathematical problems using a computer.

Alpha (α) is the level of statistical significance selected for a test by the investigator. It is usually 0.05 or one chance in 20, signifying that a difference of the observed magnitude could occur by chance, in one out of 20 such sets of data. Alpha is also used for a number of other purposes in statistical analysis, such as in *linear regression* where it represents the intercept with the y-axis when x = zero. Those other uses are clear in context. (See *significance level* and *Type I error*.)

Ambient means "surrounding", as in "ambient concentrations in the workplace were x..." meaning concentrations in the air. Recently, the word has often been used in a redundant fashion, and the best remedy is to delete it ("... in the ambient environment ...").

Analysis of covariance (ANCOVA) is a technique for evaluating data produced by an experimental design which has both continuous and discrete independent variables. The variable of most interest is assessed for significant differences, by statistically holding the other variable constant. An example could be a simultaneous regression of percent survival (the effect) on toxicant concentration (the continuous variable), for two species of daphnid (the discrete variable). If the primary interest was the relationship of effect to concentration, ANCOVA could be used to assess this by holding the effect of species constant.

Analysis of variance (ANOVA) is a formal mathematical procedure for determining whether a significant difference exists among the means or variances of samples which arise from different *treatments*. Common treatments would be exposures to different toxicant concentrations including a control, or location in different regions of a plume such as control, nearfield, and farfield. In ANOVA, the "background" variances within samples are used to tell whether or not an overall difference exists among the treatments, but this cannot tell which one(s) differ from which others. Accordingly, ANOVA is often used before *a multiple-comparison test*. (See Section 7.4.)

Angle or *angular transformation* is frequently used to refer to the *arcsine transformation*.

ANCOVA (see *analysis of covariance*).

ANOVA (see analysis of variance).

Arcsine transformation can be applied to data that are proportions or percentages, which tend to form a binomial distribution. The purpose would be to make the variances consistent and the distribution nearly normal, so that parametric statistical analyses could be used. The transformation is the arcsine of the proportion in question. Arcsine is also abbreviated to arcsin. Its value for any proportion can be obtained from many software programs and scientific calculators, or looked up in a table given in most statistic texts. The transformation was useful before the advent of modern computational aids that ease the burden of manual calculations. Use of arcsine transformations could well be avoided nowadays.

Asymptotic (see *threshold*).

Beta (**ß**) is the probability of making a Type II error (concluding a "false negative", that there is no significant difference when one is actually present). Beta is related to the *power* of a test. The symbol Beta is also used as a population *parameter* in the formula for a regression, in which the symbol **ß** represents the slope. (See also *Type 2 error* and *linear regression*.)

Bias occurs when estimates differ in a predictable manner from their true (but unknown) value. For example, poor water quality could bias the results of toxicity tests towards greater apparent toxicity. (See *accuracy* and *precision*.)

Binary is equivalent to *quantal*. Binary information is "either-or"; an observation on an individual *experimental unit* must take one of two possible forms. A seed germinates or does not germinate, etc.

Binomial distribution or *binary probability distribution* describes the likelihood that a *binomial random variable* is represented by some specified value. It may be thought of as a curve showing the pattern of frequencies associated with the proportions for a positive quantal event (e.g., mortality in a toxicity test). The frequencies depend on the number of observations and the chance (probability, p) that the event will occur. For sample sizes that are moderate (say 25) or greater, with $p \approx 0.5$, the binomial distribution resembles the familiar bell-shaped *normal distribution*. In such a distribution, many of the observations would cluster near the proportion 0.5, with fewer and fewer observations as proportions diverged further towards zero or 1.0. (See also, *probability distribution*.)

Binomial variable or *binomial random variable* is a count of the number of individuals possessing one of the two possible *quantal/binary* characteristics (e.g., death) in an experiment.

Bioassay is a test in which living organisms are used to estimate the strength or potency of a material such as a medical drug. In pharmacology, the potency is usually estimated by comparing with results for a standard preparation, tested simultaneously. "Bioassay" has also been applied to environmental tests, but *toxicity test* identifies such tests and their objectives more specifically, and is the recommended term.

Block is a sub-set (or all) of the experimental treatments. Each block is subjected to the same experimental treatments. For example, a set of tests in one growth chamber could represent a block, with the purpose of removing one source of variability, namely the possibility of different ancillary conditions within the set of tests, as the result of conditions in different chambers. Blocking receives little emphasis in toxicity tests of Environment Canada, because the procedures for testing are closely described, i.e., there is emphasis on reducing extraneous variation through experimental design and control of the test apparatus and setup. (See also *replicate*.)

Chemical is any element, compound, formulation, or mixture of a substance that might be mixed with, deposited in, or found in association with soil, sediment, or water.

Chi-square or χ^2 is a test statistic that is sometimes used in assessing the fit of a model to a set of data.

Chronic means occurring during a relatively long period of exposure, usually a substantial proportion of the life span of the organism, such as 10% or more. The word is often distorted in environmental toxicity, to signify "sublethal" or sometimes "life-cycle", but such incorrect use should not occur. "Chronic" should be used in its standard sense as in other fields of toxicology, and the proper specific terms ("sublethal" etc.) should be used in other situations.

Chronic toxicity refers to the long-term effects of a poison that are related to changes in basic processes such as metabolism, growth, or reproduction. Chronic effect might, however, be assessed by mortality or length of life.

Coding means conversion of original measurements, to numbers or symbols that have some advantage for subsequent analysis. Coding could use a simple arithmetic operation to produce values that are easier to work with. For example, a series 842, 846, 849, 845 ... could have 840 subtracted from each term, yielding 2, 6, 9, 5 ... In that example, a calculated mean value would have the same characteristic of being 840 low, compared to the original data. Coding could also be done to represent categories, for example, females might be coded as 1 and males coded as 2.

Coefficient of determination. See R^2.

Coefficient of variation (CV) is the standard deviation divided by the mean, usually expressed as a percentage.

Collinearity refers to correlation between independent variables. *Multicollinearity* has the same meaning. If two independent or explanatory variables are strongly correlated, the second variable brings little additional information to explain the effect. Strong collinearity can inflate the variance of partial regression coefficients. If it is severe, it can prevent the matrix inversion that is required for estimation of parameters. Collinearity might be detected by (1) creating a correlation matrix of independent variables and examining it for strong correlations, or (2) examining the signs and magnitude of regression coefficients to ensure that they make sense. (See also, *linear regression*.)

Concentration-effect. See *dose-response.*

Confidence limits are so similar in magnitude to *fiducial limits* that they are treated as the same thing in this document. These limits on an EC50 or ICp represent upper and lower concentrations, within which the true endpoint is thought to lie, for a stated level of probability. The 95% confidence limits represent a statement that there is a 19 out of 20 chance that the true endpoint falls within those specified limits.

Confound means that an undesired variable is influencing the experimental results in a non-random manner. For example, if all replicates of given concentrations were placed together in a regular and sequential pattern, within the array of test containers, then location within the laboratory would be confounded with the concentration being tested.

Contaminant refers to a physico-chemical or biological material that has been added to a natural substrate such as air, water, soil, or sediment, as the result of some direct or indirect human activity. It is detectable through testing, and might produce a chemical or physical change in the substrate, but might not cause adverse biological effect. The term is usually applied to materials that are present in low concentrations, in a situation where adverse biological effects have not been demonstrated. Various agencies use the term in particular senses, and it has specific meanings under some national and international definitions or regulations. *Contamination* can refer to the process by which the material is added, or the status of the substrate or biota, in having the foreign material present.

Control is a sample in an investigation that duplicates all the factors that might affect results, except the specific condition or treatment being studied. In toxicity tests, the control must duplicate all conditions in the treatment exposure but must contain no test material (i.e., no toxicant). The control is used as a check for apparent toxicity resulting from basic conditions such as quality of dilution water or health and handling of organisms. *Control* is synonymous with *negative control*. (See also *positive control*, *salinity control*, *solvent control*, *control sediment*, *reference sediment*, and *reference soil*.)

Control sediment is clean sediment, which could be taken from an uncontaminated site, or could be formulated (reconstituted). For cultured organisms, it could be a sample of sediment identical to that used for the culture. This sediment must contain no added test substance and must enable an acceptable rate of survival or performance of the test organisms, as specified in the method. (Contrast with *reference sediment*.)

Convergence is the tendency of a series of numbers to move towards a definite limit or common point.

Correlation means that the magnitude of one variable tends to change proportionally with the magnitude of another variable. One variable does not necessarily cause the change in the other. (See also *regression*.)

Correlation coefficient is properly called *multiple correlation coefficient*. (See *R*.)

Criteria are defined by CCREM (1987) as scientific data, evaluated to derive the recommended limits of water quality for particular uses. The singular is *criterion*. See also *quality guideline*. More common usage, in the USA and elsewhere, gives *criterion* the meaning that is assigned to *guideline* in this glossary. For example, Rand (1995) defines a water quality criterion as "an estimate, based on scientific judgments, of the concentration of a chemical or other constituent in water which, if not exceeded, will protect an organism, an organism community, or a prescribed water use or quality with an adequate degree of safety". These descriptions for water apply equally to other substrates such as soil.

Crossed refers to an experimental design in which all possible combinations of factors exist. For example, with two factors, gender of the test organism and concentration of toxicant, and a measured effect of toxicant residue in the tissues, it is possible to design an experiment in which each gender is exposed to each concentration. (See *nested*.)

Datum is a numerical fact, observation, or item of numerical information. The plural is *data*.

Degrees of freedom is a characteristic of a given set of data that is being analyzed statistically. It is a statistical concept that refers to the freedom with which a value can be specified. For example with "n" observations and a fixed mean, any values may be chosen for "n-1" observations. The last observation, however, is fixed by the mean and the values of the first n-1 observations. The degrees of freedom are n-1. Degrees of freedom are often used when estimating an "average variance" or mean square error.

Dependent variable (see *variable*.)

Derivative (see *partial derivative*.)

Discrete variable (see *variable*.)

Distribution refers to the way in which a particular characteristic is spread over members of a class, often represented graphically by a curve. As commonly used, distribution is synonymous with *probability distribution*. It is the relative frequency for the values that a variable can have. For example, in the daphnid reproductive test, an average number of neonates is usually in the range of 18 to 22 per adult. The relative frequency of values in that range is much higher than for a value of, say, 35. The probability distribution describes these relative frequencies. It can be used to determine how probable it is for an observation, or range of observations, to occur, for a given distribution.

Dose is the amount of a chemical or toxicant that has entered a test organism. The dose is unknown for most tests of environmental toxicity, which assess the effect of concentrations in the medium. See also *dose-response*.

Dose-response is an adjectival expression, used to refer to classical concepts of pharmacology or toxicology such as "dose-response relationships", the distribution of observed changes in organisms as related to the amount of drug or toxicant. The expression is used in very general ways in environmental toxicology, although *concentration-effect* would usually be more appropriate. As mentioned previously, most tests of environmental toxicity deal with ambient concentrations, rather than the doses within the organisms. Similarly, the word "response" is suitable for use in medicine or pharmacology, where the human or other organism can show an improvement from a dose of a curative drug, while in toxicology, the organism is not so much responding to the toxicant, as suffering an effect of it.

EC50 is the *median effective concentration*. It is the concentration of material in water (e.g., mg/L), soil or sediment (e.g., mg/kg) that is estimated to cause a specified toxic effect to 50% of the test organisms. In most instances the EC50 and its 95% confidence limits are statistically derived by analyzing the percentages of organisms showing the specified effect at various test concentrations, after a fixed period of exposure. The duration of exposure must be specified (e.g., 72-h EC50). The EC50 describes quantal effects, lethal or sublethal, and is not applicable to quantitative effects (see *ICp*). Other percentages could be used, see *ECp*.

ECp has the same meaning as *EC50*, except that "p" can represent any percentage, and is to be specified for any particular test or circumstance. Some investigators and agencies, particularly European and international, have mistakenly used ECp to mean ICp, but the distinction is important and should be maintained.

Ecotoxicology has the same general meaning as *environmental toxicology*.

ED50 is the *median effective dose*. The meaning is similar to *EC50* except that it refers to a toxic *dose*.

Effect, in toxicology, means a measurable biological change. The change could be structural, physiological, behavioural, etc. In a toxicity test, the biological change should be assessed against a background of measurements on the organisms in the control. The statistical analysis generally considers the degrees of effect that are beyond the control measurements, and are therefore presumed to result from exposure to toxic components of the material being tested.

Effective concentration, see *EC50*.

Effluent is any liquid waste (e.g., industrial, municipal) discharged to the environment. There is no need to use the expression "whole effluent".

Elutriate is an aqueous solution obtained after adding water to a solid material (e.g., soil, sediment, tailings, drilling mud, dredge spoil), shaking the mixture, then reclaiming the liquid by centrifuging it, filtering it, or decanting the supernatant.

Endpoint is the *statistic* that is estimated as the result of a test. The endpoint is used to characterize the results of the test (e.g., the *ICp* or the *LC50*). It is not recommended to also use this term to mean the effect on an organism, or the variable being observed, such as size of an organism at the end of the test, although that usage might be encountered (OECD, 2004).

Environmental toxicology has the same general definition as *toxicology*, since it is a branch of that science. However, the focus is on effects towards wild living organisms and natural communities, without excluding the safety of humans as part of the ecosystems.

Error. *Pairwise error rate* or *comparison-wise error rate* is the ratio of the number of incorrect inferences, to the total number of inferences made. *Experiment-wise error rate* is the probability of encountering at least one *Type I* error during the course of making all the comparisons (for a given effect) in the experiment. For example, in the context of a survey of toxicity in sediments, the comparisons would be between the mean effect for each of the locations with the mean for the control. The "experiment" would be the entire survey. This error rate would not include comparisons of any other biological effects. (See also *Type I* and *Type II* errors.)

Experiment-wise error (see *Error*).

Experimental error (see *precision*).

Experimental unit is the smallest independent unit or element in a toxicity test, to which a *treatment* is applied. The experimental unit shows an effect which is measured and becomes a *datum*. An example is one container of organisms in a toxicity test. (The organisms within the container would be *sampling units*.) If there were two or more containers exposed to one treatment, each container would be an experimental unit and it would also be a *replicate*. (See also *sampling unit* and *block*.)

Exponent, in mathematics, is the superscripted symbol denoting the number of times that a quantity is to be multiplied by itself, as $5^2 = 5 \times 5 = 25$. (See also *logarithm*.)

Fiducial limits (see *confidence limits*).

Field replicate (see *replicate* and *replicate samples*).

Flow-through describes tests in which solutions in test vessels are renewed continuously by the constant inflow of a fresh solution, or by a frequent intermittent inflow.

General linear model (GLM) does not refer to a specific mathematical technique, but describes a class of models with similar characteristics and approach. There is a single dependent variable (possibly with multiple measurements on an experimental unit) which is a function of an independent variable or variables. The GLM framework includes simple linear regression, analysis of variance, analysis of covariance, repeated measures, and others.

Generalized linear model (*GLIM, GLiM,* or *generalized linear interactive model)* is a further generalization of the approach used for the *general linear model*. This unified approach estimates the parameters of models in which the effect is normally distributed, and also when effects belong to any member of the exponential family of distributions, including binomial, logistic, Poisson, and log-normal. In toxicology, a researcher could use GLIMs to assess the dependence of a quantal or quantitative effect on a single independent variable such as

concentration (by regression), or a more complicated structure of independent variables such as group treatment (ANOVA), or treatments and covariates (ANCOVA). The GLIM is not well defined nor circumscribed for understanding by non-statisticians.

Geometric series, or geometric progression, signifies that each successive number in the series is greater than the preceding number by a factor that is constant through the series (e.g., 3, 6, 12, 24 ...). The numbers are also in a *logarithmic* series.

Geometric mean is a measure of central tendency for a set of observations. It can be useful because it is less influenced by extreme values than is the more familiar arithmetic mean. For n values in a set, the geometric mean is the n^{th} root of the product of all the values (i.e., multiplied). It can also be calculated as the antilogarithm of the arithmetic mean of the logarithms of the values.

GLIM (see *generalized linear model*).

GLM (see *general linear model*).

Gompertz distribution (see Weibull.)

Good Laboratory Practice (GLP) is a set of standards governing the experimental design, collection of data, and conduct of scientific and technical studies in the laboratory. The Standards Council of Canada and Environment Canada (EC) have GLP programs. Standards are also promulgated by the Organisation for Economic Co-operation and Development (OECD) and the United States Environmental Protection Agency (USEPA).

Goodness of fit is a statistical statement or index of how well observations conform to a theoretical or estimated distribution. Measurement of chi-square is the usual index, and is used as the example here. Chi-square measures how well the observed frequencies fit the theoretical frequencies. The degree of fit (the "goodness") is expressed by the numerical value of chi-square. [Zar (1999) points out that "poorness of fit" might have been a better name, since higher and higher values of chi-square indicate increasingly worse agreement of observations with the theoretical pattern.] A value of zero for an index would indicate a perfect fit, and a value of infinity could theoretically result from a bad enough fit, but an index cannot have a negative value.

Graded effect (see *quantitative*).

Guideline (see Quality guideline).

Heteroscedasticity refers herein to data showing heterogeneity of the residuals within a scatter plot (see Figures O.2B and O.2C in Appendix O). This term applies when the variability of the residuals changes significantly with that of the independent variables (i.e., the test concentrations or treatment levels). When performing statistical analyses and assessing residuals (e.g., using Levine's test), for test data demonstrating heteroscedasticity (i.e., non-homogeneity of residuals), there is a significant difference in the variance of residuals across concentrations or treatment levels. (See also *homoscedasticity* and *residuals*.)

Homoscedasticity refers herein to data showing homogeneity of the residuals within a scatter plot (see Figure O.2A in Appendix O). This term applies when the variability of the residuals does not change significantly with that of the independent variables (i.e., the test concentrations or treatment levels). When performing statistical analyses and assessing residuals (e.g., using Levine's test), for test data demonstrating homoscedasticity (i.e., homogeneity of residuals), there is no significant difference in the variance of residuals across concentrations or treatment levels. (See also *heteroscedasticity* and *residuals*.)

Hormesis is an effect in which low concentrations of the test material act as a stimulant for performance of the test organisms compared to the control organisms ("better" than the control). At higher concentrations, deleterious effects are seen. A more general category of "low-dose stimulation" would include other possible causes of stimulation, such as solvent effects, experimental error, or "sufficient challenge" among laboratory organisms.

ICp is the *inhibiting concentration for a (specified) percent effect*. It represents a point estimate of a concentration of test material that is estimated to cause a designated percent impairment in a quantitative biological function such as the size attained by organisms during a growth period. For example, an IC25 for weight would be the concentration estimated to result in organisms having a dry weight 25% lower than that attained by control organisms. This term should be used for any toxicological test which measures a quantitative effect or change in rate, such as attained size, reproductive performance, or respiration. The term EC50 is incorrect for these quantitative tests (see *median effective concentration*). The ICp may be estimated by regression or, if necessary, by the procedure of *smoothing and interpolation* using the computer program ICPIN.

Incipient as in *incipient LC50* or *incipient EC50* for acute quantal effects, is that stimulus intensity (i.e., concentration) at which an effect can be expected in (just) 50% of the test organisms after indefinitely long exposure. The rationale is that this represents the concentration that would just be sufficient to affect the median organism (the "typical" or "average" organism). The original, more general, and still useful term is *incipient lethal level* (Fry, 1947). Equivalent terms are *threshold EC50*, *time-independent EC50*, and *asymptotic EC50*, all referring to the *toxicity curve* becoming parallel to the time axis. The term incipient has roots in environmental physiology and avoids conflicting connotations of the word "threshold". The definition of "incipient" becomes more arbitrary and difficult for sublethal quantitative effects, which lack the obvious and customary criterion of median effect, as used for quantal tests. For quantitative tests, incipient might best be defined as the lowest concentration at which there was a significant deleterious change in the effect being assessed (such as growth). In practice, such an estimate for a quantitative effect would vary with the design and precision of the test.

Independent variable (see *variable*).

Interquartile range (see *quartile*).

Iteration is a mathematical process used to estimate the parameters of a regression (i.e., to "fit a line"). It involves successive approximations to the estimates by cycles of calculation, each cycle building on the previous approximation and improving the estimates.

Laboratory replicates (see *replicate* and *replicate samples*).

Leachate is water or wastewater that has percolated through soil or solid waste.

Least squares is a method of fitting a line to a set of data. It minimizes the sum of squares of deviations of the observed values from the respective predicted values.

Lethal means causing death by direct action. Death is usually defined as the cessation of all visible signs of movement or other activity, and failure to show such signs upon gentle external stimulation.

Lethal concentration, see *LC50*.

LC50 is the *median lethal concentration*, i.e., the concentration of material in water, soil, or sediment that is estimated to be lethal to 50% of the test organisms. The LC50 and its 95% confidence limits are usually

derived by statistical analysis of percent mortalities in several test concentrations, after a fixed period of exposure. The duration of exposure must be specified (e.g., 48-h LC50). Other percentages could be specified, such as LC20.

LD50 is the *median effective dose*. Definition as *LC50* except that it is expressed in terms of the *dose*.

Life-cycle test is one in which the organisms are observed from a life stage in one generation to at least the same life stage in the next generation.

Linear regression is a statistical procedure for estimating the parameters of a model that describes the relationship between an effect or response (the *dependent variable*) and a set of explanatory variables [the *independent variable*(s)]. "Linear" does not refer to the shape of the line but to the nature of the equation describing it. Linear models are relatively simple, in that their parameters (*a, b,* etc.) can be estimated by evaluating a single formula. The phrase "simple linear regression" is often used when only one explanatory variable is used. A simple linear model would be the familiar equation for a line, $\boldsymbol{Y = a + bX}$ in which Y is the dependent variable, X is the independent variable, and **a** and **b** are parameters. However, linear regression can include curved lines as well as straight ones, for example a quadratic model could be included ($Y = a + bX + cX^2$). Statisticians use the term "linear" to describe models in which the *partial derivatives* of the model with respect to a parameter are independent of any other parameters. See also *partial derivative, regression,* and *nonlinear regression*.

LOEC is the *lowest-observed-effect concentration*. This is the lowest tested concentration of a material which has an effect that is different from the control, according to the statistical test used for analysis. See also, NOEC. (The O does not signify "observable", a grammatical mistake sometimes seen. The LOEC is associated with an effect that the investigator actually noted *(observed)*. An effect at an even lower concentration might have been *observable*, given a more powerful experiment, more time spent in scrutinizing the organisms, a better microscope, etc. Nor should the word "harmful" be incorporated into the name of the endpoint (NOHEC). It should be left to the experimenter to designate the kind of effect that is included, without imposing some outside definition of "harmful".)

Log is an abbreviation for *logarithm* to the base 10.

Logarithm is a method of mathematical coding. Here, logarithm ("common logarithm" or "log") is the power to which a fixed "base" of 10 must be raised, to produce the number represented by the logarithm. Thus a logarithm of 2 would represent $10^2 = 100$, i.e., log to the base 10 of 100 = 2, or $\log_{10} 100 = 2$. Some examples provide insight: $\log_{10} 700 = 2.84510$, $\log_{10} 70 = 1.84510$, $\log_{10} 7 = 0.84510$, $\log_{10} 0.7 = -0.15410$ (or 9.84510 - 10). Adding (subtracting) logarithms is equivalent to multiplying (dividing) the numbers they represent. See also *exponent*; in $10^2 = 100$, the exponent is 2, thus related to logarithms. *Natural* or napierian logarithms, (or *ln* as in ln 100 = 4.60517) use the base "e", which has a value of 2.71828... Either type of logarithm can be used in toxicological work as long as the usage is consistent throughout a set of calculations. The base "e" is important in mathematical concepts such as compound interest, exponential function, probability theory, growth equations, etc.

Logarithmic, as in logarithmic series, signifies that the logarithm of each number in a series is greater by a constant amount, than the logarithm of the preceding number. The numbers themselves could also be said to be in a *geometric series*, since each would be greater than the preceding by a constant multiplier.

Logit is the logistic equivalent deviate. It is a specific transformation of data that can be applied to the proportions of organisms affected in a quantal toxicity test, usually resulting in a straightening of the sigmoid curve of effect. To obtain logits, the proportion of organisms affected (p) at a given concentration is divided

by (*1 - p*). The logarithm of the result is taken, and that is the logit. See also Section 4.5.1 and *probit*, with further discussion in Appendix J. Logits also provide a useful way of fitting a regression to quantitative data. Results are formulated as the proportions of organisms which attained specified values of the measured effect. Examples are given in Section 6.5.8 with further details in Appendix O.

Logistic distribution is a statistical distribution function which has been found to be of value in quantal assays and in regressions of quantitative data. (See *logit*.)

Lowest-observed-effect concentration (see *LOEC*).

LT50 is the *median lethal time*, i.e., the exposure time that is estimated to be lethal to 50% of the test organisms for a given concentration of test material. Successive observations of mortality in each of a series of concentrations can allow an estimate of LT50 for each concentration, sometimes advantageous in providing a more revealing *toxicity curve*. The usual statistical techniques for LC50 are not valid for LT50.

MATC is the *maximum acceptable toxicant concentration*. It has been defined in various disparate ways but is now generally considered to be synonymous with *TOEC*; the latter term is recommended here.

Matrix is used, particularly in work with sediment and soils, to refer to the background physical and chemical nature of a sample. Accordingly, *matrix effect* refers to the action of these background characteristics on test organisms. This is intended to refer only to the background effects, without those caused by any contaminants present.

Material is the sum of all the *substances* which it contains. A material has more or less uniform characteristics. Soil, sediment, or surface water are materials.

Maximum likelihood estimation is a mathematical method for obtaining estimates of parameters in a relationship of interest. *Maximum likelihood estimates (MLE)* attempt to estimate the values of the parameters that would result in the highest likelihood of observing the data actually collected (SPSS, 1996). For example, the parameters might be the mean and variance of a distribution of data. "The *likelihood* that a set of parameters should have any particular values is defined to be a quantity proportional to the probability that, if these be the parameters, the totality of observations should be the data recorded" (Finney, 1978, p. 58). This is not the same concept as least squares, nor minimum chi-square.

Mean or *arithmetic mean* is the most widely used measure of central tendency in a set of data. It is the sum of all the observations, divided by the number of observations. Because it considers the numerical value of each observation, the mean can be thought of as the "centre of gravity" for a set of data.

Median is the middle measurement in a set of data that has been ordered from small to large or large to small. It is the number of items in the series that is being divided, not the arithmetic values of those items. If there were an odd number of items in the series, the median would be the middle item. If there were an even number of items, the median would usually be the average of the numerical values of the two middle items. If it were more complicated, such that more than two items at or near the middle had the same numerical value, the median would be interpolated under an assumption that the middle values were evenly ranged across the middle interval; statistics textbooks provide proper formulae. The median expresses less information than the mean because it does not take into account the actual value of each measurement. However, it can be a good choice to describe central tendency in a skewed population, because extremely high (or low) measurements will not affect the median as much as the mean (Zar, 1999). See also *quartile*.

Median effective concentration/dose (see *EC50/ED50).*

Median lethal concentration/time/dose (see *LC50/LT50/LD50*).

Medium, in toxicity testing, is the material that surrounds or carries the organisms. Examples include the culture medium for bacteria (the nutritive broth or substrate), the water in which fish are swimming, or the soil surrounding earthworms. The plural is *media* ("The bacteria were cultured in a medium that ..." or "We tested several media to determine...").

Minimum Significant Difference (MSD) is the magnitude of difference in measured results that would have to exist between the control and a test concentration, in order to conclude that there was a significant effect at that concentration, according to the statistical test being used.

MLE (see *maximum likelihood estimation*).

Monotonic is a sequence of numbers in which each value is either (a) greater than or equal to the preceding one, or (b) less than or equal to the preceding one.

Multicollinearity. (See *collinearity*.)

Multiple-comparison test is a statistical procedure which can be used to distinguish how mean effects differ statistically from each other in an experiment that has more than two treatments. Sometimes called multiple-range tests. (See Section 7.5.)

Multiple correlation coefficient. (See *R*.)

Multiple regression is a relationship in which the magnitude of a dependent variable is governed by two or more independent variables. For example, $\mathbf{Y = \alpha + ß_1X + ß_2X + ß_3X}$. (See also *polynomial*.)

Natural logarithm (see *logarithm*).

NEC is the *No-effect concentration*, the level of a toxicant that is thought to have no effect whatsoever on a specified organism. The NEC is somewhat of an idealized concept, and must be predicted or estimated by modelling or extrapolation. It is analogous to a *parameter* of a *population*. From the results of any given toxicity test, the NEC must be deduced rather than observed, because more tests or different kinds of tests might reveal effects at lower concentrations.

Negative control has the same meaning as *control*.

Nested refers to an experimental design in which all possible combinations of a factor cannot exist (compare *crossed*). If a test involves the factors gender of the organism and concentration of toxicant, with triplicate measurements of toxicant residue in the tissues, it is not possible to design an experiment in which each animal is found at each combination of factors. The triplicate measurements of residue are subsamples and are nested within the factor "animal".

NOEC is the *no-observed-effect concentration*. This concentration is the next lower from the *LOEC*, among those concentrations tested. (Almost always, the NOEC is also the highest tested concentration whose effect on test organisms is not different from the control, according to the statistical test used for analysis. It is possible, however, that irregular response could result in no significant effect at a concentration higher than the *LOEC*. That situation is avoided by the definition given for NOEC.)

Nonlinear regression is similar to *linear regression*, but the *partial derivatives* of a parameter are not independent of other parameters. The term does not refer to the shape of the line showing this relationship. The dependent variable *cannot* be expressed as a linear combination of parameter values multiplied by values of the independent variable (SPSS, 1996). The formula describing the regression might be multiplicative, e.g.,

$Y = \alpha\beta^X$ which is the formula for exponential growth (Zar, 1999). An iterative approach is required to estimate the model parameters. (See also *linear regression* and *regression*.)

Nonparametric analysis is a statistical technique that does not assume any underlying distribution for the data. The technique does not make use of the *parameters* (such as mean and variance) of the *population* from which the samples were drawn. Nonparametric testing draws inferences about the population but not about the parameters of the population. (See also *parametric analysis*.)

No-observed-effect concentration (see *NOEC*.)

Normal distribution (or *normal probability distribution*) is a symmetric bell-shaped array of observations. The array relates frequency of occurrence to the magnitude of the item being measured. In a normal distribution, most observations will cluster near the mean value, with progressively fewer observations toward the extremes of the range of values. The shape is determined by the mean and standard deviation, with 68.3%, 95.4%, and 99.7% of the observations included within plus or minus one, two, and three standard deviations of the mean, respectively. Not all bell-shaped curves are normal, and normality is defined by a particular and complex equation which includes the mean and standard deviation, and also the constants π (3.14159) and e (the base of natural logarithms). The normal distribution plays a central role in statistical theory because of its mathematical properties. It is also central in biological sciences because many biological phenomena follow the pattern. Many statistical tests assume that data are normally distributed, and therefore it can be necessary to test whether that is true for a given set of data.

Normal equivalent deviate (NED) is the standard deviation from the mean of a normal distribution, associated with a particular probability. In other words it is a unit of divergence from the mean of a normal distribution, expressed in terms of the standard deviation of that distribution. One NED is one standard deviation out from the mean. A *probit* is simply an NED, with 5 added to avoid negative values on one side of the distribution.

Observation. (See *variable*.)

One-tailed test is a statistical test designed and appropriate for the situation in which the investigator is interested in whether a variable differs in only one direction, from the base of comparison (e.g., is the variable greater than the base?). In a *Two-tailed test*, the investigator wishes to determine whether the variable differs in either direction from the base of comparison, i.e., is it significantly different?

Outlier is an extreme observation, a measurement that does not seem to fit the other values from a test.

Pairwise error (see *Error*.)

Parameter is a word with several connotations. In mathematics it is a property or characteristic of a *population*, such as the mean or the median. For any one population, the parameter has a value that is constant. If a sample were taken from the population, the mean or median of that sample would not be a parameter, but would be called a *statistic*. The statistic would almost certainly vary among different samples taken from the same population. A toxicity test uses a sample of test organisms, so the *endpoint* is a statistic, which is regarded as an estimate of the true value (the parameter for the total population of organisms). In biostatistics, it is a convention to use Greek letters to represent population parameters, and Latin letters to represent sample statistics. In common usage, parameter has somewhat uncertain meanings, but usually signifies limits, boundaries, guidelines, or restrictions (Burchfield, 1996). "Parameter" is often misused, even in government publications, in places where the correct word would be *variable*. A very common misuse is in lists or tables of chemical measurements, in which the chemicals being measured are said to be "parameters"; they are not, they are variables, and investigators should avoid that mistake.

Parametric analysis uses a biostatistical method that considers the *parameters* of the *population* from which the samples were drawn. Usually, this means that if two sets of samples are being compared, the two populations from which they were drawn must have normal distributions and equal variances. The samples being subjected to the analysis must conform to the characteristics assumed for the population. (See also *nonparametric analysis*.)

Partial derivative has reference to the independent variables of a function. A *derivative* can be explained in terms of a very simple function such as Y = aX. The "Y" is the dependent *variable*, "X" is the independent *variable* and "a" is a *parameter*. The derivative is the change in Y relative to the change in X (i.e., the slope). Hence the derivative is $\delta Y/\delta X = a$. Now if the function has two or more independent variables, the derivative must be taken with respect to each of them, in order to describe the slope. For example, if the function is Y = aX1 + bX2, there are two *partial derivatives*, namely $\delta Y/\delta X1 = a$, and $\delta Y/\delta X2 = b$.

Partial effect means that some of the test organisms in a container showed the effect, and some did not. This could be applied to lethal effects, as in *partial mortality,* which would mean that some organisms died and some did not.

Partial mortality (see *partial effect*.)

Point estimate is a single numerical value that has been calculated or judged to represent a set of toxicity data, e.g., EC50 or IC25.

Poisson distribution is one involving counts of random occurrence of an item, either in space or time. An example would be counts of algal cells in the squares of a grid. If the probability was small (but constant), and the number of observations large, the Poisson would be similar to the *binomial* distribution.

Polynomial refers to an equation of a *multiple regression* in which some of the terms have exponents. For example, $\mathbf{Y = \alpha + ß_1X + ß_2X^2 + ß_3X^3}$.

Pollutant is the substance, material, or form of energy which causes *pollution*, or is capable of causing pollution if discharged to the environment in sufficient quantities. (See *pollution* and *contaminant*.)

Pollution is the addition of a material or a form of energy such as heat, to some component of the environment, in such an amount as to cause a detectable change which is deleterious to some organism or to some human use of the environment. Some regional, national, and international agencies have formal definitions of pollution, which should be honoured in the appropriate contexts.

Population, in mathematics, is the collection of all possible values of a variable (such as the lengths of all individual fish in a lake). Or the word could represent all the individuals in the group of interest (such as the fish in a lake). *Universe* is used synonymously in mathematics. (See also *sample*.)

Porewater is the water occupying space between sediment particles. The amount of porewater is expressed as percent of the wet sediment, by weight.

Positive control is a toxicity test with a reference toxicant, used to assess the sensitivity of the organisms at the time that a test material is evaluated, and also to assess the precision of results obtained by the laboratory for the reference chemical.

Power is, loosely, the probability of correctly concluding that there is a difference between the variables being tested. Formally, it is "the probability of rejecting the null hypothesis when it is in fact false and should be rejected". In effect, it is the opposite of making a *Type II error*, in which an investigator accepts the null

hypothesis when there is actually a difference. The probability of making that Type II error is called ***ß***, and *power* is represented by (1 - **ß**). Power cannot be directly and precisely set by the investigator, before doing a toxicity test. Power can be increased, however, by strengthening the toxicity test (more organisms, more replicates, etc.). Calculating power at the end of a test is rather complex, but power is related to *Minimum Significant Difference*, which can be estimated by standard procedures in many statistical tests which operate on quantitative data.

Precision is the closeness of repeated measurements to each other, and is often assessed by the variance or standard deviation. A group of measurements could be very precise, but have poor *accuracy*. Measurements might be both precise and accurate but still have *bias*. If measurements are made on each of several organisms *(sampling units)* from each of two or more containers *(experimental units)* at a given concentration *(treatment)*, the variation among containers is the *experimental error* and determines the precision of the mean value of measurements at that concentration. The variation among measurements of individual organisms within a container is *sampling error*. In quantal tests, the proportion affected in a container is the observation on the experimental unit, while the effect shown by an individual organism (affected or not affected) is an observation on a sampling unit; the precision is a function of the number of organisms in a container. In the usual methods of analysis for quantal tests, the data from replicates are pooled, so the variation within concentrations cannot be used directly.

Probability is the likelihood of an event, measured by the ratio of the favourable cases to the whole number of cases possible.

Probability distribution is a function describing the probability that a random variable is equal to or less than some unspecified value. A familiar example is the bell-shaped normal distribution. If the random variable is equal to 1.645, the probability of being less than 1.645 is 95%. (See also *distribution*.)

Probit is a unit of divergence from the mean of a normal distribution, expressed in terms of the standard deviation of the distribution. It is the *normal equivalent deviate*, with 5.0 added to avoid the confusion of negative values on one side of the distribution (a step that is convenient for understanding, but not actually necessary in these days of computer analysis). The practical use of probits, in estimating an LC50 or EC50, is to straighten the sigmoid curve of the accumulated normal distribution, which shows percent effect as a function of log concentration. (See also *probit scale*.)

Probit scale has a central value of 5.0, representing the expected median effect in a quantal toxicity test (an expected 50% of the organisms would be affected). For most practical purposes, the scale from 3 to 7 probits would be adequate. Probit 2 would represent an expected effect on 0.1% of test organisms, probit 3 would be 2.3% of organisms, probit 4 would be 16%, probit 6 = 84%, probit 7 = 97.7%, and probit 8 = 99.9%. (See *probit*, *normal equivalent deviate*, and Section 4.5.1.)

Probit regression (often called probit analysis) measures the relationship between the strength of a stimulus and the proportion of cases that exhibit a selected effect caused by the stimulus (after SPSS, 1996). Regression would normally use *maximum likelihood estimation* or iterative reweighted least squares, to estimate the ECp and the relationship of probit of effect to logarithm of concentration. The effect being analyzed is *quantal*.

Protocol is an explicit set of procedures for a test or experiment, formally agreed upon by the parties involved, and described precisely in a written document.

Pseudoreplicate is a false *replicate*. A common example in toxicity testing would be mistakenly calling the organisms within a test vessel “replicates”. Using pseudoreplicates as replicates in a statistical test would be a gross error.

Quadratic refers to a type of equation for a regression, which contains a third parameter and X^2.

Quality Assurance (QA) is a program within a laboratory, intended to provide precise and accurate results in scientific and technical work. It includes selection of proper procedures, sample collection, selection of limits, evaluation of data, *quality control*, and qualifications and training of personnel.

Quality Control (QC) consists of specific actions within the program of quality assurance. It includes standardization, calibration, replication, control samples, and statistical estimates of limits for the data.

Quality guideline is a scientifically based numerical concentration limit or narrative statement recommended to support and maintain a designated use of a medium such as soil, air, or water ("soil quality guideline", etc.). A *quality objective* has the same definition as guideline except that it applies to a specific site. Some provinces have established lists of *water quality objectives*, and they reflect "officially desired conditions". A *quality standard* is an objective that is recognized in enforceable environmental control laws or regulations, promulgated by a government.

Quantal is an adjective as in quantal data, quantal test, etc. A *quantal effect* is one for which each test organism either shows the effect of interest or does not show it. For example, an animal might either live or die, or it might either develop normally or abnormally. Such data usually fit a *binomial distribution*. The term *dichotomous* means the same, is now more frequent in statistical literature, and is more easily understood. (See also *binary*, *binomial variable*, *discrete*, and *quantitative*.)

Quantitative is an adjective, as in quantitative data, quantitative test, etc. A *quantitative effect* is one in which the measured effect can take any whole or fractional value on a numerical scale. An example would be the weights attained by individual organisms at the end of a test. Such data usually fit a normal distribution. *Continuous* can be a synonym and is commonly used by statisticians concerned with toxicology, especially in Europe. *Graded* was used to mean the same thing in this context, by the early giants of toxicology (Gaddum, 1953), but is no longer considered an appropriate term. (See also *quantal*.)

Quartile identifies one of the three values in a ranked series of numbers, which divide the series into four equal parts. It is the number of items in the series that is being divided, not the arithmetic values of those items. One-fourth of all the ranked numbers in the series would occur before the *first quartile*, and three-quarters would occur after it. Three-fourths of the numbers would occur, in the series, before the *third quartile*, and one-fourth after it. The second quartile is called the *median*, and half of the items in the ranked series come before it, and half after it (see *median*). The *interquartile range* is the absolute value of the difference between first and third quartiles. Usually it is easy enough to pick the quartiles and median by inspecting the series. However, in short series it can be questionable to decide on quartiles which divide the series appropriately, and various sources differ on precise definitions and methods of calculation (see Appendix R).

R is the *multiple correlation coefficient.* It is the square root of the *coefficient of determination* (see R^2). It estimates the multiple correlation coefficient (ρ or *rho*) in the population that was sampled. R is also equal to the Pearson product moment correlation (usually designated as *r*) between the predicted and observed values in a regression analysis (see *linear regression).*

R^2 is the *coefficient of determination* or *coefficient of multiple determination*, often referred to as the "R^2 value. It is the ratio of the sums of squares accounted for by a regression model, to the total sums of squares about the mean. In a regression context, the coefficient of determination measures the proportion of variability in the effect measured, that is explained by the regression model (see also *R*).

Random sample is a selection of individuals (or items or elements) from a *population*, in which each individual has an equal probability of being included in the selection. For most statistical techniques, random sampling is required to make valid inferences.

Range is the difference between the highest and lowest values in a set of data. It is usually called "the range" and is often given as the actual high and low values.

Receiving water is surface water (e.g., in a river, lake, or bay) that has received a discharged waste, or else is about to receive such a waste (e.g., it is in a flowing river just upstream from the discharge point). Description must be provided to indicate the meaning intended.

Reference method is a procedure for testing toxicity, which has an explicit set of instructions and conditions described precisely in a written document. Unlike other multi-purpose (generic) biological test methods published by Environment Canada, the use of a reference method is usually triggered by the testing requirements of specific regulations.

Reference sediment is a field-collected sample of presumably clean sediment that has properties (e.g., particle size, compactness, total organic content) closely matching those of the sample(s) of test sediment except for the degree of chemical contamination. It is often selected from a site uninfluenced by the source(s) of contamination but within the general vicinity of the survey stations where samples of test sediment are collected. It is used to describe *matrix* effects in the test, and may also be used as a control and as a diluent to prepare concentrations of the test sediment. (See also, *control sediment.*)

Reference soil is a field-collected sample of presumably clean soil that has properties (e.g., texture, structure, pH, organic content) as similar as possible to those of the sample(s) of test soil, except that it is free from the chemical contamination being assessed. It is often selected from a site uninfluenced by the source(s) of contamination, but within the general vicinity of test samples, and thus might be subject to pollutional influences other than the one(s) being studied. It is used to describe *matrix* effects in the test, and may also be used as a control and as a diluent to prepare concentrations of the test soil.

Reference toxicant is a standard chemical used to measure the sensitivity of the test organisms, and to assist in establishing validity of the toxicity data obtained for a test material. In most instances, a toxicity test with a reference toxicant is performed to assess (a) the sensitivity of the organisms at the time the test material is evaluated, and (b) the precision of results obtained by the laboratory over a period of time which includes several or many tests of that reference toxicant.

Regression, as a statistical technique, determines the relationship between two or more variables. The term refers to the activity of estimating, or to the relationship after it has been calculated. The magnitude of a *dependent variable* (such as size) is a function of the magnitude of another variable or variables, the *independent variable(s)* (such as concentration). The reverse is not true. This can be called *simple regression* if there are only two variables. (See also *linear regression*, *nonlinear regression* and *correlation*.)

Repeated measure refers to making more than one numerical observation over time, on the same experimental unit. "Repeated measures analysis" is a separate category of statistical methods for these types of observations, a category not covered in this document.

Replicate is a repetition of a *treatment (= experimental unit)*. (See also *block* and *replication.*)

Replication is the repetition of sets of treatments in groups. (See also *replicate*.) A replicate (as a noun) is a single test chamber containing a prescribed number of organisms *(= sampling units)*, either in one concentration *(= treatment)* of test material, or in a control. In a toxicity test with five test concentrations and a control, using three replicates, 18 test chambers would be used, i.e., three chambers for each treatment. A replicate must be an independent test unit, and therefore the test material in a chamber must not have a connection to the test material in another chamber. Any transfer of organisms or test material from one replicate to another would invalidate a statistical analysis based on the replication. As a verb, to *replicate* is to repeat a *treatment* or *experimental unit*. Experimental error (the random variation among the experimental

units) is estimated from the replicates. (See also *block*, *replicate samples*, *experimental unit*, *pseudoreplicate*, *sampling unit,* and *treatment*.)

Replicate samples are separate samples of soil, sediment, etc., collected in the field, using identical methods, and at the same sampling station. By definition each replicate is subject to the same treatment. The purpose is to provide a more representative appraisal of the quality of the sampled substrate, and to allow the variation in that quality and/or the variation in sampling the substrate to be estimated. The replicate samples must be stored in separate containers. The replicate samples might be used to set up replicates within each treatment in a toxicity test; that is often recommended in tests with soils or sediments. These replicates in the test would be true *field replicates*, so the test would assess variation in the test material and in sampling it, as well as any variation between replicates created by conditions in the laboratory. *Laboratory replicates* in a test would be two or more replicates for each treatment in a test, created by splitting or taking a *subsample* of the sample of test material. Such a test would merely indicate variation due to conditions in the laboratory, and must not be construed to indicate the variation in the test material (say, sediment in a lake) or in sampling that material.

Laboratory replicates are usually pointless in a toxicity test, and are not recommended unless for convenience in size of containers or some similar reason. They could, however, be of some use in regression, for differentiating between the error of measuring an effect, and the actual deviation of an effect from the fitted line. For chemical analysis, laboratory replicates could be taken to assess *precision* of chemical measurements.

Residuals, in regressions, are the differences between each observed value and the value predicted for it by the equation.

Resistance is a characteristic of an organism, describing its ability to delay the manifestation of designated effects of a toxicant or other environmental identity for a period of time which is a function of the level of the identity. Ultimately, the organism will succumb (after Fry, 1947). (Contrast with *tolerance*.)

Response, in this document, is considered a synonym of *effect*. The latter term is preferred in toxicology, because damage to the test organism by the toxic substance is not so much a case of the organism responding, as it is a consequence of the toxicant's action. Although the term *dose-response* is often used in a general way to describe relationships in toxicity tests or *bioassays*, "concentration-effect" would usually be more specific in environmental toxicology. In any case, the effect or response is almost always the dependent or "y" variable in a statistical model.

"Safe" concentration is that concentration of the test substance estimated to allow normal life history and reproduction of organisms within their natural habitat. The "safe" concentration is a biological concept, not a statistical endpoint from an experiment, and is usually given quotation marks to indicate the uncertainty about whether it is completely safe. (See also *NEC*.)

Salinity has traditionally referred to measurements of the total mass of dissolved salts in a given mass of solution, described in terms of g/kg or "parts per thousand" (‰). Today it is empirically measured from standard relationships of density or conductivity, and results are unitless (APHA *et al.*, 1992).

Salinity control is a separate control chamber or set of chambers in a toxicity test with marine organisms. It serves the purpose of a normal *control*, and also assesses any effect of less-than-optimal salinities in the test chambers. The term would not be relevant to tests in which all treatments were adjusted to a standard optimal salinity; such tests would simply have a "control" with the same salinity as test concentrations. (That is the case for marine tests under the Environmental Effects Monitoring Program of Environment Canada, which also has special controls related to the technique used for adjusting salinity; see Section 2.7). If salinity adjustment was not done in the test concentrations, there should be a control at favourable salinity, and in addition, an extra set of *salinity controls* duplicating the test salinities. The purpose would be to indicate deleterious effects of low (or high) salinity acting alone. However, if there was some deleterious interaction of the divergent salinity with toxicity of the test material, the extra set of salinity controls would not indicate that interaction.

Sample is a part of a population, selected by an investigator. Usually, the intention is to use information from the sample to make inferences about a *population*. Accordingly, it is important to clearly define the population of interest and to obtain a representative sample from that population; this is often done by *random sampling*.

Sampling error. (See *precision*.)

Sampling unit is an observational unit within an *experimental unit*. An example would be one organism in a container of organisms exposed to a given *treatment*. (See also *replicate*.)

Sediment is a natural particulate material that has been transported to, and deposited at, the bottom of a body of water. The term can also describe a substrate that has been experimentally prepared.

Serial dilution is a series of test concentrations in which each lower concentration differs from the preceding one by a constant factor (dilution), such as 100, 50, 25, 12.5%. Such a series can be obtained by successive dilutions of a given waste, stock solution, or stock material.

Sigma ($\sum$) is mostly used to mean "the sum of ...". Lower-case sigma (σ) is mostly used to signify the standard deviation of a population.

Significance or *significant* refers, in this document, to differences between or among groups, that cannot be ascribed to chance alone. The distinction is made on the basis of a formal statistical test. Unless otherwise stated, a 5% level of probability is assumed, i.e., the difference would not be expected to occur by chance more than 5% of the time, if the experiment or test were repeated many times.

Significance level is defined statistically as the probability of rejecting the null hypothesis when it is true. In other words, it is the probability of erroneously concluding that a treatment (such as a toxicant concentration) had a significant effect, when in fact, it did not. Toxicity investigators might also use the words as follows: "... there is a difference at a significance level of 5%". (See also *Type I error* and *Power*.)

Simple regression (see regression).

Skew means asymmetry. Here, it refers to asymmetry in a plotted frequency curve for a given distribution of data. The plot of a classical normal curve is symmetrical, that is the left and right sides of the curve will be mirror images about the mean, and the median will have the same value as the mean. In a curve that is skewed to the right, the curve looks asymmetric with the right tail stretched out, and the mean is higher than the median. If that curve is cumulated, it will be noticeable that the upper part of the curve is stretched out to the right in a sweeping curve. (See Section 9.1 and Appendix H.1.)

Soil elutriate (see *elutriate*).

Solvent control is a special type of control that might be necessary in a toxicity test, most likely an aquatic test. It is appropriate for any toxicity test in which a solvent is used to obtain the desired concentrations of a poorly soluble chemical that is being tested. A solvent control must be run simultaneously with the standard control(s). Usually the solvent control must duplicate the conditions in the standard control, except that it must contain the highest concentration of solvent that is found elsewhere in the test. For a satisfactory result, the performance of organisms in the solvent control must not be "worse" than performance in the regular control. (See Section 2.7.2.)

Spiking refers to the addition of a known amount of chemical or substance to a soil or sediment. The substance is usually added to a clean or control soil/sediment, but sometimes to a contaminated one. The substance added would usually be a single chemical, but might be a test soil/sediment. After the addition, the soil/sediment is mixed thoroughly.

Standard is an enforceable, defined level of quality (see *Quality guideline*).

Standard deviation describes the divergence of individual observations in a sample, from the mean value for that sample. It is the square root of the *variance*, and by definition, it can only be a positive number. The symbol *"SD*" is used in this document, in keeping with common practice of biologists and some other recognition (Zar, 1974); mathematicians would use the symbol "*s*".

Standard error is the usual abbreviation for the *standard error of the mean*. It is represented by the symbol "*SE"* (or *"s"* with subscript x-bar, as commonly used by mathematicians). Standard error can be calculated for *any* statistic, for example the estimate of a slope in a linear regression has a standard error. However, the most common use in toxicology is for the SE of a sample mean. The SE of a sample mean is calculated as the *standard deviation* of the sample, divided by the square root of the number of observations in the sample. That calculated SE is an estimate of the divergence that would be shown among a number of mean values, if those means represented a number of samples taken from the same population. The standard error for a group of means, therefore, is the equivalent of the standard deviation for a group of observations in a single sample. In practice, the standard error is estimated from a single sample, as indicated previously.

Static describes aquatic toxicity tests in which test solutions are not renewed during the test.

Static-renewal describes aquatic toxicity tests in which test solutions are renewed (replaced) periodically during the test, usually at the beginning of each 24-h period of testing. Synonymous terms are "renewal", "batch replacement" and "semi-static".

Statistic is a quantity or measurement that characterizes some property of a *sample*. (See *population*).

Sublethal means detrimental to the organism, but below the level that directly causes death within the test period.

Subsample is part of a single sample. In statistical terms, subsamples are multiple observations of a given characteristic on one experimental unit. A subsample must represent a single time of collection. If collected over time, the observations would fall in the category of *repeated measures*.

Substance is a particular kind of material having uniform properties; often the term would apply to a chemical compound.

Threshold as in *threshold EC50*, see *incipient EC50*.

TOEC is the *threshold-observed-effect concentration*. Its true value lies somewhere between the NOEC and LOEC; it is estimated as the geometric mean of those two concentrations, for the convenience of having a single endpoint.

Tolerance is a characteristic of an organism, and in environmental toxicology it means the ability to withstand specified levels of an environmental identity for an unlimited time. It was originally defined in lethal temperature work with fish, using the description "the *zone of tolerance* in which the animal will never die from the effects of that particular identity alone" (Fry, 1947). (See also *tolerance distribution* and *resistance*.)

Tolerance distribution is used by statisticians to mean the pattern *(distribution)* of effects among organisms exposed to a single concentration of a toxic agent. The usage can be illustrated by an example of growth shown in a group of organisms exposed to a given concentration of toxicant. Individual organisms will exhibit a range of effects. There will be a mean growth effect, with individuals showing some dispersion around that mean. That distribution around the mean is the *tolerance distribution*. If another group is exposed to a different concentration, the mean effect will change, but it is assumed that the tolerance distribution will

remain the same, i.e., the same variance. The statistical use differs from the established definition of *tolerance* in biology and toxicity work. "Tolerance distribution" has been avoided in this document.

Toxic is an adjective or adverb meaning that a chemical, substance, or material is present at a location, in sufficient quantity to cause adverse effects on living organisms, or that the material could fulfil that role.

Toxicant is a chemical, substance, or material that can cause adverse effects on living organisms (i.e., a poison).

Toxicity is the inherent potential or capacity of a material to cause adverse effects on living organisms.

Toxicity curve is a graph of successive endpoints of a test or tests, with concentration plotted against time, both on logarithmic scales (e.g., log LC50 versus log exposure-time). The curve can indicate whether a threshold of toxicity was reached during the test, i.e., a time-independent asymptote of concentration, an important item of knowledge for any toxic chemical (see *incipient LC50*). A toxicity curve is usually for lethal effects, since for most sublethal tests, definitive observations of effect are available only at the end of the test.

Toxicity test is a determination of the effect of a material on a group of selected organisms under defined conditions. An environmental toxicity test usually measures either (a) the proportions of organisms affected (*quantal*) or (b) the degree of effect shown (*quantitative*) after exposure to specific concentrations of chemical, effluent, elutriate, leachate, receiving water, sediment, or soil.

Toxicology in its broad sense, is the science that defines limits of safety of chemical agents. There is no limitation on the scientific disciplines that may be used, on whether the tools are in the laboratory or field, or whether the studies are at the molecular or ecosystem level. The scientific studies must, however, be designed with a goal of defining limits of safety. (See also *environmental toxicology*.)

Toxic unit (TU) is an expression of the toxic potency of a waste material, or of a substance contained in a medium such as soil, sediment, water, or air. Potency of the waste material or substance is expressed as a multiple of (= fraction of) a standard endpoint of toxicity. Toxic units of a waste material such as an effluent would be calculated as 100% (i.e., the strength of the effluent) divided by the endpoint as a percentage (e.g., an effluent with an LC50 of 10% would have 100/10 = 10 lethal toxic units). For a toxic substance contained in a substrate or medium, the example of a chemical dissolved in water may be taken. Its *lethal toxic units* would be calculated as the actual concentration of the chemical in the water, divided by the LC50 of that chemical. *Sublethal toxic units* would be calculated by using a defined sublethal endpoint (such as IC25) as the denominator. For example, if a chemical was present in water at 5 mg/L and the IC25 of that chemical was 10 mg/L, there would be 5/10 = 0.5 sublethal TU present, i.e., half of the sublethal effect-level. Toxic units are unitless in terms of chemical concentration. They are conceptually convenient since their numerical value increases as the potency increases.

Toxin is a poisonous substance, especially a protein, produced by living cells or organisms and capable of causing disease or other deleterious effects when introduced into an organism. A toxin is also capable of stimulating production of an antitoxin. An example would be paralytic shellfish poison produced by marine dinoflagellates ("red tide"). News media and careless environmental activists have almost ruined the meaning of this word by using it for all kinds of toxicants.

Treatment is, in general, an intervention or procedure whose effect on a *sampling unit* is to be measured. More specifically, in toxicity testing, it is a condition or procedure applied to the test organisms by an investigator, with the intention of measuring the effects on those organisms. Usually, the treatment would be a concentration of a potentially toxic material. The treatment might include several containers at the same concentration, each of them an *experimental unit* and also a *replicate*. In the testing of sediments and soils,

treatment means a specific material being tested (e.g., site sediment, site soil, reference soil, or negative control soil) from a particular sampling station. (See also *sampling unit*.)

Two-tailed test (see *one-tailed test*).

Type I error, commonly designated as ***α*** *(alpha)*, occurs when an investigator rejects a null hypothesis that is true. In other words, the investigator concludes that there is a significant difference, but actually, there is none.

Type II error, commonly designated as ***β*** *(beta)* occurs when an investigator fails to reject the null hypothesis when it is false (concludes that there is no significant difference, but actually there is).

(A) *Variable* is a characteristic which differs from individual to individual, case to case, or observation to observation. Thus, the variable is a characteristic of the individuals or cases in a *population* of individuals or cases. A variable could be the concentration of a chemical, the height of plants, the number of progeny, or similar items. The value measured or recorded for the variable is an *observation*. The variable can be *continuous*, taking any value of a spectrum within the possible range (such as concentration of a chemical or the weight of a midge larva). Or, the variable can be *discrete*, signifying that it can take on any positive or negative values such as 0, 1, 2, 3, e.g., the number of leaves on a plant. The two designations correspond, respectively, to *quantitative* and *quantal* data. An *independent variable*, in an analysis, would be the one that is fixed, usually by the investigator, and is being used to predict the corresponding value of the *dependent variable*. The value of the latter is governed by the choice of the independent variable. In a toxicity test, concentrations would be the independent variable, and effect would be the dependent variable. (See also *binomial variable* and *parameter*.)

Variance is a measurement which describes the divergence of individual observations in a sample from the mean value for that sample. It is calculated by (a) subtracting the mean from each observation, (b) squaring each result from step (a), (c) adding together the values from step (b), and (d) dividing the result of step (c) by one less than the number of observations. The symbol for variance is s^2. [The variance for the theoretical population from which the sample was drawn would use the symbol σ^2, and would be estimated from a sample by the method described, except step (d) would divide by the number of observations.] The units of a variance are usually omitted; if given, they would be squares of the original units, and might not make sense. (See also *standard deviation*.)

(A) *Warning chart* is a graph used to follow changes over time, in the endpoints which measure toxicity of a *reference toxicant*. The date of the test is on the horizontal axis and the effect-concentration is plotted on the vertical logarithmic scale.

Warning limits allow an investigator to evaluate the variation in toxicity tests with a *reference toxicant*. The limits are plus and minus two standard deviations, calculated logarithmically, from the historic geometric mean of the test endpoints.

(The) *Weibull* distribution is a generalized version of an exponential model. It can be used for empirical fits to dose-effect data. The distribution is sigmoidal, but allows the shape of the curve to differ above and below the inflection point, an advantage over probit or logit distributions. The Gompertz model is essentially equivalent to the Weibull, and is useful for nonlinear regression (see Section 6.5.8).

Weighting is adding "arithmetic emphasis" to certain values in a series, so that those values have greater influence on whatever calculation is being carried out. The purpose is to compensate for some perceived irregularity or deficiency in a set of data. A particular value might be weighted to indicate that it deserved greater emphasis because it was based on a large sample, or represented a group of observations with a small variance.

Whole sediment is the entire intact sediment that has had minimal manipulation following collection or formulation. It is not a form or derivative of the sediment such as an elutriate or a resuspended sediment.

Whole soil is the entire intact soil that has had minimal manipulation following collection or formulation. It is not a form or derivative of the soil such as an elutriate or a leachate.

Acknowledgements

This document was written by John B. Sprague (Sprague Associates Ltd., Salt Spring Island, B.C.) with the direct technical input of Barry A. Zajdlik (Zajdlik & Associates Inc., Rockwood, Ont.) and based on suggestions from Glenn F. Atkinson (Atkinson Statistical, Calgary, Alta.). The document is derived from previous guidance documents, reports, and the published literature on toxicology and statistics, supplemented by suggestions from scientific and technical people in government, industry, and academia from across Canada and elsewhere. Richard P. Scroggins (Chief, Biological Methods Division, Environment Canada, Ottawa, Ont.) was Scientific Authority for the project, providing technical assistance and guidance throughout the work. Stella Wheatley (Polaris Scientific and Technical Editing, Ottawa, Ont.) edited and formatted the document and prepared some of the figures. The descriptions of models in Section 6.5.8 and the instructions for use of SYSTAT in Appendix O were compiled and standardized by Juliska Princz (EC, Ottawa, Ont.) from recent methods documents published by Environment Canada.

Invaluable assistance is gratefully acknowledged, from the following people who provided comments and suggestions on drafts at one or more stages of development: Larry W. Ausley (North Carolina Dept. of Environment, Raleigh, N.C.); Uwe Borgmann (National Water Res. Inst., Burlington, Ont.); Kenneth G. Doe (EC, Moncton, N.B); Natalie Feisthauer (Stantec Consulting, Guelph, Ont.); Hector F. Galicia (Springborn Smithers Laboratories (Europe) AG, Horn, Switzerland); John W. Green (DuPont, Newark, Del.); Christine S. Hartless (USEPA, Washington, D.C.); Janet McCann (Univ. of Waterloo, Waterloo, Ont.); Donald J. McLeay, (McLeay Environmental Ltd., Victoria, B.C.); Cathy McPherson (EVS Environment Consultants, North Vancouver, B.C.); Jennifer Miller (Miller Environmental Sciences Inc., Innisfil, Ont.); Mary Moody (Saskatchewan Research Council, Saskatoon, Sask.); Serge Morissette (Ministère de l'Environnement, Sainte-Foy, Que.); Marion Nipper (Texas A & M Univ., Corpus Christi, Tex.); Niels Nyholm (Technical Univ. Denmark, Lyngby, Denmark); R. Jeanette O'Hara Hines (Univ. of Waterloo, Waterloo, Ont.); Juliska Princz (EC, Ottawa, Ont.); Hans Toni Ratte (Rheinisch-Westfalishce Technische Hochschule, Aachen, Germany); Jim Reid (ESG International, Guelph, Ont.); Julie E. Schroeder (Ont. Min. Environment and Energy, Etobicoke, Ont.); and Wout Slob (Nat. Inst. Public Health & Environment, Bilthoven, The Netherlands).

Thanks are extended to the following people who provided computer analyses, information, reports, other major tangible help, or organized, synopsized, or participated in workshops on this project: Howard Bailey (EVS Environment Consultants, North Vancouver, B.C.); Joy Bruno (EC, North Vancouver, B.C.); Craig Buday (EC, North Vancouver, B.C.); Curtis Eickhoff (BC Research Inc., Vancouver, B.C.); Paula Jackman (EC, Moncton, N.B.); Nicky Koper (Univ. Alberta, Edmonton, Alta.); Nancy Kruper (EC, Edmonton, Alta.); Don Larson (IRC Consultants, Richmond, B.C.); Michelle Linssen (EC, North Vancouver, B.C.); Tim Moran (Pollutech Enviroquatics, Pt. Edward, Ont.); David Moul (EC, North Vancouver, B.C.); Michael D. Paine (Paine, Ledge and Associates, North Vancouver, B.C.); Janet Pickard (BC Research Inc., Vancouver, B.C.); Linda Porebski (EC, Gatineau, Que.); Danielle Rodrigue (EC, Ottawa, Ont.); Gladys L. Stephenson (ESG International Inc., Guelph, Ont.); Armando Tang (EVS Environmental Consultants, North Vancouver, B.C.); Becky-Jo Unis (Hydroqual Laboratories, Calgary, Alta.); and Graham van Aggelen (EC, North Vancouver, B.C.).

Introduction

Toxicity tests are powerful tools for investigating and resolving problems of environmental contamination and pollution. However, data from the tests must be analyzed properly to obtain valid estimates of endpoints. This document is intended to assist in making proper choices for statistical analysis of tests in environmental toxicology. In particular, it is intended for use with more than 20 toxicity test methods published by Environment Canada, for microbes, aquatic and terrestrial plants and invertebrates, and fish (EC, 1990a–c; 1992a–f; 1997a,b; 1998a,b; 1999b; 2000a,b; 2001a; 2002a; 2004a–c; see listing in Appendix A). This document focuses on mathematical and statistical methods for analysis of results; another guidance document deals with general approaches and interpretations in environmental toxicology (EC, 1999a).

1.1 Purposes and Objectives of this Document

Key Guidance

- *The primary goal of this document is to help establish good statistical practices at Canadian laboratories which do toxicity testing within programs of Environment Canada.*
- *Statistical tests in current use are discussed, with indications of preferred methods, and others which hold promise. Some examples are worked.*
- *Explanations are aimed primarily at new laboratory personnel. The focus is on standard testing rather than research projects.*
- *Advice is provided on recognizing and dealing with "difficult" types of data. Some common mistakes are explained.*

Simply expressed, this document is intended to offer information in three areas.

(a) Additional guidance to users of Environment Canada's single-species toxicity tests. The orientation is aimed at new laboratory personnel, rather than experienced investigators.

(b) Some explanation of the statistical reasons behind the procedures in toxicity tests. This document, however, does not teach basic statistics.

(c) Comments on existing statistical tests and some profitable approaches that might become available in the future.

The basic objectives of this document were defined by a Statistical Advisory Committee and other interested parties who met following the 20th Annual Aquatic Toxicity Workshop in Quebec City in 1993 (Miller *et al.*, 1993). These objectives are to provide:

(1) guidance on the statistical methods for biological tests, thereby leading to a more standardized approach for calculating toxicity test endpoints;

(2) background information on the characteristics, strengths, and weaknesses of various statistical procedures, and on the importance of their assumptions;

(3) methods for assessing whether the results of an experiment provide definitive answers to the initial questions that were posed;

(4) examples of applying statistical methods and interpreting their results; and

(5) guidance on recognizing and dealing with "difficult" data.

Background papers were presented at the meeting in Quebec, and discussion of nine topics resulted in the specific recommendations acted upon in relevant parts of this document. Objectives (3) and (4) are dealt with as relevant throughout the document, when particular procedures are described.

The glossary has been extended into examples and explanations allowing investigators to relate their general knowledge to the particular applications in toxicity testing for environmental purposes.

Some examples accompany the information in the document, along with mention of pitfalls and deficiencies. Experienced investigators might feel that there is too much emphasis on common mistakes, but this is important. Some data collected in national regulatory programs have shown relatively low-grade but common mistakes in procedure. Perhaps this results from a lack of knowledgeable or experienced laboratory personnel. There is no need for the next group of personnel to repeat mundane mistakes.

Accordingly, this document is largely concerned with current, established methods in environmental toxicology, and does not attempt to plough new ground. Although it points to new and advanced techniques which appear to be in the forefront of methods development, it can only give limited coverage of their developing analytical procedures. Some of the new methods will no doubt become advantageous standards, while others will fall by the wayside. Generally, such new methods are being proposed by specialists, or teams which already have good statistical input. Similarly, it seems that Canadian studies of complex local problems of toxicity and pollution generally receive the direct expert statistical advice that they need.

This document is not designed to provide advice for programs of basic research. Researchers, and those applying advanced statistical techniques, might find useful guidance in Sections 4 and 6 herein, in an international document (OECD, 2004), and in the references listed in the following paragraphs.

General guidance in statistics may be obtained from textbooks such as Snedecor and Cochran (1980), Steel *et al.* (1997), Zar (1999), and Wardlaw (1985), the last reference being written in a style that is particularly friendly to biologists. Statistical background directly related to toxicology and other environmental studies can be obtained from Newman (1995), Gad (1999), Manly (2000), and Millard and Neerchal (2000). More specialized and classical toxicological topics are described by Finney (1971; 1978), Ashton (1972), Hewlett and Plackett (1979), or Hubert (1992). Collett (1991) and Fleiss (1981) offer advice on analysis of binary data and proportions, the basis of quantal effects. Broader philosophical education on major ideas of applied statistics is provided in a pleasant anecdotal book called "The lady tasting tea ..." (Salsburg, 2001). Some other compendia with apparently relevant titles might not be of immediate assistance in statistical analysis of tests (OECD, 1995; Grothe *et al.*, 1996). Finally, there are tremendous sources of information (and sometimes misinformation), at the ever-changing web sites on the Internet. Some are useful for general concepts or particular statistical techniques. Some excellent sites are equivalent to chapters in textbooks or to lecture notes.

1.2 How to Use this Document

A person carrying out a toxicity test would usually start with a test method described in a document published by Environment Canada or another organization. Each document tells the statistical methods to be used, with descriptions that are adequate for most purposes.

If further explanation is desired, an investigator could consult one of Sections 3 to 8 herein, dealing with different types of tests and analyses. Those sections start at the level of an individual method document, and provide additional advice on analysis, avoiding pitfalls, and some of the reasons behind a choice of methods. The specific test procedures described in each method document published by Environment Canada are definitive, and should be followed in programs of Environment Canada. The present document does not supersede any of the specific test methods.

New investigators might wish to scan Sections 2, 9, and 10. Section 2 deals with aspects of test design, some common mistakes, and offers background information. Section 9 is more general, provides some statistical background, and covers methods of testing for differences. Section 10 gives advice on certain difficult kinds of results, including outliers and stimulation at low concentrations of toxicant. Readers can check out "key guidance" at the start of sub-sections, to decide which parts might be useful. As mentioned, the glossary is detailed, to provide additional orientation.

Appendices constitute a second part of the document. They provide some technical or detailed background for assertions in the main part of the document, for readers wishing such coverage.

1.3 Main Categories of Tests

Key Guidance

- *Two primary categories are: (a) single-concentration tests to compare a material with a control or reference material; and (b) multi-concentration tests intended to estimate an ECp, ICp, or NOEC.*
- *The multi-concentration category may be subdivided into: (a) quantal variables (each organism either reacts or fails to react); or (b) quantitative or continuous variables (e.g., weight of individuals).*
- *Dual-effect tests often have quantal and quantitative measurements, which at this time, are best analyzed separately.*

A primary division in types of tests, is between those using a single concentration of the test material, and those using several concentrations.

Single-concentration tests compare toxic effects in a sample to the performance in a control (or to a reference material or some other special sample, location, or condition). For example, a single sample of sediment might be compared to a reference sediment. These "single-concentration tests" are shown on the left of Figure 1. Variations might include just one test sample, or a number of samples from different locations, tested simultaneously with a control or reference. They might or might not involve replication. These types of tests are described in Section 3.

Multi-concentration tests use several fixed concentrations and a control, designed to estimate an ECp, ICp, or NOEC. This kind of test is shown in the middle and on the right of Figure 1, which provides a general overview of the types of tests.

Both single- and multi-concentration tests might make observations of a quantal effect or a quantitative one (Figure 1). In quantal tests, a direct count of exposed organisms classifies them as either affected or not, i.e., *binary* or dichotomous data. Results are best fitted to a binomial distribution and analyzed by statistical techniques appropriate to such a distribution (e.g., the chi-square test). However, most quantal tests in environmental toxicology are lethal tests. Analysis is customarily by probit or logit regression or a substitute method. The usual endpoint is the *median lethal concentration* (LC50), or *median effective concentration* (EC50), a more general term that also covers sublethal effects. Quantal tests are described in Section 4.

Quantitative tests measure some continuously variable effect such as size of the organism. Historically, they were once called "graded tests" (Gaddum, 1953), a term which no longer seems appropriate. The data provided by such tests can be called "continuous" data. The usual endpoint is the *Inhibiting Concentration for a specified percent decrease in performance* (ICp). For example, the IC25 could represent 25% lower weight of organisms than in the control groups. The pattern of results often follows the familiar normal curve (actually a cumulated normal curve). Ideally, endpoints would be estimated by regression, and methods for regression have recently been prescribed in some standard tests published by Environment Canada (Section 6.5.8). The alternative but less favoured approach is to use hypothesis testing to estimate the NOEC and LOEC, as described in Section 7.

Some tests have dual effects and they usually involve a quantal effect such as mortality, and a quantitative effect such as weight or reproduction. At this time, results of dual-effect tests should be analyzed separately (central stream of Figure 1), because a technique has not been developed for handling the correlated effects. The analysis showing the lower effect-concentration is usually adopted as the endpoint for the test. Dual-effect tests are described in Section 8.

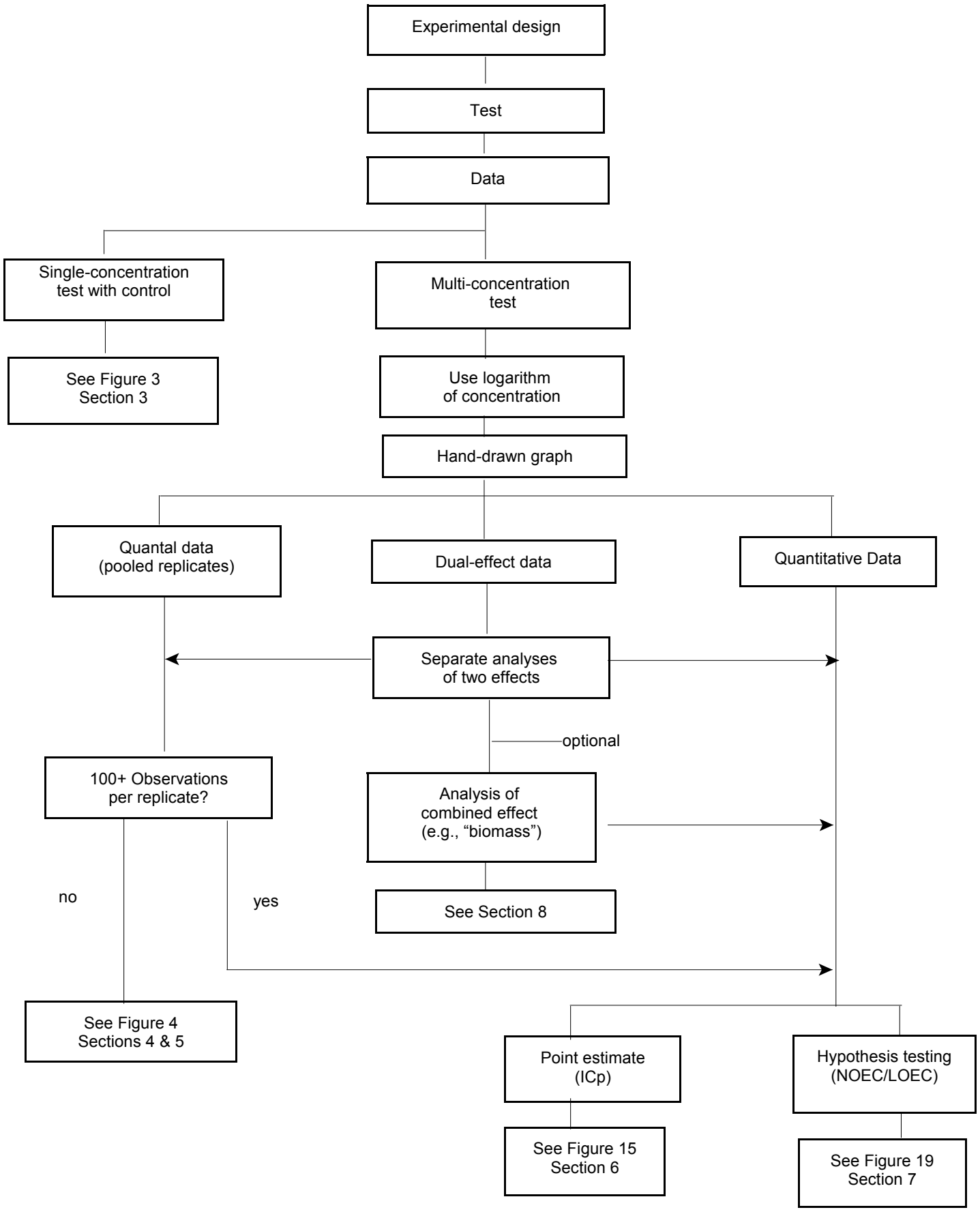


Figure 1 A flow-sheet of the main categories of environmental toxicity tests covered in this document.

General Design and Analysis

The best statistical procedures cannot remedy a poorly designed experiment. Environment Canada's *Biological Test Methods* include advice on design, and it should be followed to obtain data that are valid for statistical analysis. In particular, investigators should never disregard instructions on randomization, replication, or controls (see Sections 2.4, 2.5, and 2.7). Other important components of design are covered in Sections 2.1, 2.2, and 2.3.

The variability of results can sometimes be reduced by judicious use of more than the minimum specified numbers of organisms, replicates, or concentrations. Such improvement would be especially desirable in tests that will be used for registration of new chemicals or in legal proceedings.

2.1 Participation of a Statistician

It is a truism that a statistician should be involved at all stages of a test, including the design, analysis, and statement of findings. Often that is easier said than done, especially for small laboratories or commercial testing; however, the principle remains. In these days of modern communication, it should be possible to work out some system of consultation that is rapid and economical. The remedy might be occasional broad sessions of advice on desirable approaches and remedies for situations likely to be encountered. In part, the present document intends to help orient investigators, to gain more benefit when a statistician is contacted.

Sometimes the advice from a statistician might simply be to proceed with a standard toxicity test. The design and methods of analysis should be agreed on, and the statistician might alert the investigator to potential difficulties. If there is any reason to expect an irregular distribution of data, that should be considered at the design stage of the test. If an effect is likely among control organisms, that also needs to be considered during design.

At the same time, the investigator who might be a biologist, should keep in mind her/his priorities, and balance these with the advice from a statistician. If asked about narrowing the limits of error, a statistician can scarcely avoid recommending more test concentrations, more organisms, etc. The investigator must make difficult decisions to balance those recommendations with practical matters of cost, time, facilities, and work priorities. (For the mathematically minded, Bayesian mathematics allows the notions of probability and cost to be combined, to determine the cost-effectiveness of obtaining additional information. An introduction to the topic is found in Morissette (2002); it focuses on sampling contaminated sediments but is of wider relevance.)

2.2 Selecting Concentrations

Key Guidance

- *Choice of concentrations is an important and difficult aspect of test design. If results could be foreseen, an ideal design would have several concentrations causing a middle range of effects, and others spread equally above and below with a span from negligible to strong effects.*
- *The most common problem is choosing concentrations that are too close together. All of them might turn out to be either too high or else too low, spoiling the test.*
- *Choosing widely separated concentrations runs a risk of failing to obtain "partial" or middle-range effects, but this is less serious.*
- *Useful remedies would be a preliminary range-finding test, and/or a relatively large number of concentrations in the definitive test.*
- *For any choice of concentrations, good design calls for a consistent geometric series.*

In a multi-concentration test, the choice of appropriate concentrations is the most important aspect of test design. Inappropriate choice is the most common cause of "difficult" sets of results. The most frequent mistake is choosing concentrations that are too close together. An investigator might guess (wrongly) at the anticipated endpoint, and make an inappropriate and unfortunate choice of concentrations. That choice might lead to a worst possible situation, of severe effects in all concentrations, or else feeble effects in all concentrations. The test would be a failure, particularly if it had been done on a field sample that could not be duplicated. Some examples are given in Section 10.4.

To determine an ECp or an ICp, the aim should always be to have concentrations that are both above and below the endpoint. Lacking one of those ranges, extrapolation to estimate the endpoint is always undesirable, and often impossible. Unfortunately, a perfect choice of concentrations could only be made if the outcome of the test were known in advance, so an investigator is forced to use judgement. This judgement can be greatly improved by running a preliminary range-finding test, even a relatively crude one. Following that, the design of the actual toxicity test can be most improved by increasing the number of concentrations to be tested, and spacing them widely enough.

Whatever the purpose and general design of the test, it is important to select a regular geometric series of concentrations. Each concentration must show a constant proportional increase over the preceding one, for example, a two-fold increase could yield the series 4, 8, 16, 32, 64, etc. At first glance, the gap appears large between 32 and 64, but to the test organism the increase represents exactly the same doubling of chemical stimulation as between 4 and 8. Or, it represents exactly the same doubling that would occur in a series which is lower by an order of magnitude, i.e., between 3.2 and 6.4. Whichever region of the series turns out to be the important range, the same proportional increase will prevail. This helps to obtain a balanced distribution of results, and the choice is basic to all later calculations (see Section 2.3 for more detail).

2.2.1 Opposing Influences

The precision and confidence limits of a toxicity endpoint depend upon some or all of the following:

(a) the number of concentrations with "middle" or partial effects;
(b) the spread of concentrations about the endpoint;
(c) the number of replicates;
(d) the number of organisms per replicate or concentration;
(e) the variation (scatter) of data-points; and for some methods of analysis, and the slope of the regression.

The investigator can attempt to create favourable conditions for (a) to (d). Item (a) is discussed in this section.

If an ideal test design could be anticipated, it would have several concentrations in the "middle" range of effects, with an equal spread of concentrations above and below that range. In trying to select such concentrations, an investigator is pulled in two directions:

(1) make the concentrations close together in order to get a good selection of "middle" or partial effects; and

(2) spread the concentrations widely, to guarantee both low and high effects.

By far the most common problem results from the first influence. The investigator is motivated to choose concentrations that are relatively close, but the endpoint occurs elsewhere than expected. As indicated previously, that might result in failure to obtain an endpoint in a regulatory or monitoring program.

Thus, an investigator should avoid the temptation of choosing concentrations that are too closely spaced, and should pay much more attention than might be thought necessary, to the second influence listed. To inspire confidence, a set of results should include a low concentration that elicits an effect similar to the control, and a high concentration that achieves a nearly maximal effect. This design is supported by the OECD (2004), which states that the "intuitive

idea of concentrating dose levels around the ECx is not optimal. Designs that include ... sufficiently different response levels compared to the controls... perform better."

On the other hand, obeying the second influence, runs the risk of failing to obtain effects in the "middle ground", effects that bracket the defined endpoint, and govern the confidence limits. This is the lesser evil; at least the endpoint would be known to lie in a given range, which is preferable to an answer of "greater than a concentration of x" (or less than ... x). If regression is used for analysis, the more widely spread data are favourable. A regression can do little with data that encompass only part of the distribution of effects. It is important to fix the tails of the distribution -- once they are established, the middle follows because the shape of the regression was usually fixed by choosing a model.

The difficulty in choosing could usually be reduced by starting with extra groups of organisms at suitable intervals of concentration, six to eight as recommended in some methods published by Environment Canada, or even more. If necessary, fewer organisms could be used per concentration, as long as there were enough to meet minimum requirements of Environment Canada. Statistically, it is preferable to have more concentrations (partial or "middle" effects), each with fewer organisms, rather than having more organisms in fewer concentrations. As mentioned elsewhere, a case has been made that using seven fish instead of 10 at appropriate concentrations does not seriously affect the precision of the endpoint (Douglas *et al.*, 1986).

Enough additional concentrations could achieve not only a wider spread, but perhaps also the desirable small intervals between concentrations.

2.2.2 Specific Types of Tests

Quantal tests fit the pattern outlined previously. The ideal set of results would have mostly partial effects (neither 0% nor 100% effect), centred on the EC50 and bracketing it. The recommended method of analysis by probit regression has an absolute requirement for two partial effects in the series. Effects near 50% carry the most weight in estimating an EC50, and help narrow the confidence limits. However, if spreading the concentrations resulted in only one partial effect, a secondary method of analysis can be used (Section 4.5.6). Even with no partial effect, an estimate of EC50 can be made. In that case, the 0% and 100% effects would be at successive concentrations in the series, probably indicating an endpoint within reasonably narrow limits (Section 4.5.7).

Regression analysis might be used for analysis of quantitative sublethal tests. Making sure that observations spanned the region from low to high effect would once again be a very important aspect of experimental design. A fitted model would best describe three "phases": an initial non-effect or low effect, then a region of increasing effect, terminated by a region of complete effect, or a near-asymptote of little or no further change. It would be highly desirable to obtain data for each of the phases (Sections 4.4 and 4.7). For quantitative sublethal tests, some examples in Section 10.4 ("difficult results") show the uncertainties that arise if a test design does not cover a wide enough range of concentrations.

For regression analysis, it is generally more advantageous to increase the number of concentrations rather than add replicates. Although this might add to the cost of setting up a test, it might not increase the total number of organisms, and might even reduce them.

Regulatory tests usually provide less challenge in choosing suitable concentrations. In routine testing of liquid effluents, there would often be a requirement to include a 100% concentration, which would obviously fix the upper end of the series. Most interest would probably lie within the 1–100% concentration range. Accordingly, a common and adequate series in effluent tests is 100%, 50%, 25%, 12.5%, and 6.25%. Usually, the regulatory test would specify a minimum number of test organisms.

Research or investigatory tests, such as determining the toxicity of a new chemical, might require additional effort in choosing concentrations. A preliminary *range-finding test* would be an effective tactic, as long as the test materials were stable in comparison to the length of the test. This would establish profitable concentrations to be used

in a definitive test. The range-finder could be quite exploratory in its design, with only a few organisms or concentrations, and short exposure.

2.3 Logarithms of Concentration

In setting up a test, it is almost automatic to choose concentrations with a constant multiplier between successive concentrations. That provides a geometric or logarithmic series. The reasons for this default are biological, and do not have their origin in statistical considerations. This is apparently the way the test organisms "see" the concentration scale, and almost universally, this is the way that exposures are done. To do otherwise is to lose efficiency and power in a test [1]. A geometric series with any desired ratio could be used, e.g., a simple series with a ratio of 2, producing concentrations like 2, 4, 8, 16, etc. Or, dividing the desired range of concentrations by the desired number of test containers might lead to a more unusual ratio, say 1.6 (concentrations of 2, 3.2, 5.1, 8.2, etc.). Any such series with a constant multiplier would have equal logarithmic intervals. For analysis, it is customary to express the concentrations as logarithms to the base 10, but natural logarithms could be used with equal suitability, as long as there was consistency within a test.

Key Guidance

- *A geometric (= logarithmic) series of exposure concentrations is standard in toxicity tests, for good reasons. Once adopted, it remains the default, as a point of good scientific method. After statistical analysis is complete, endpoints and confidence limits are customarily converted to arithmetic values for ease of understanding. However, any subsequent mathematical treatments such as means and ANOVA should use the default logarithms unless they are demonstrated to be unsuitable.*
- *Commercial computer programs currently available for analysis of sublethal results usually violate the above principle, and they default to calculations with arithmetic values of concentration. Investigators must understand how the program behaves. In some cases the only solution is to enter the concentrations as logarithms.*
- *Endpoints calculated with arithmetic concentrations become more erroneous as the data-sets become more variable.*
- *Exposure-time is also logarithmic in nature and log time must be used for calculations.*

Most people have an intuitive understanding of the reasons for setting up tests that way, and a common-sense illustration of the rationale can be developed. Using an arithmetic series might possibly be acceptable at low concentrations (e.g., 1, 2, 3, 4, and 5 mg/L), but there would probably be little agreement to retain an arithmetic interval of unity at higher concentrations (e.g., 11, 12, 13, 14, 15 mg/L). At still higher concentrations, the interval becomes ludicrous; would anybody use 101, 102, 103, 104 ... mg/L? ... or 1001, 1002, 1003 ... mg/L? It would probably be impossible to detect a difference in effect on a test organism, between concentrations of 101 and 102, let alone 1001 and 1002. The principle of larger intervals at higher concentrations becomes self-evident, e.g., 1, 2, 4, 8 ... , or 100, 200, 400, or 1000, 2000, 4000, etc.

Thus the change in effects on test organisms is related to the proportional increase in concentration, not the absolute increase. Although an increase of 10 units from 10 to 20 mg/L is a doubling of the concentration, the same arithmetic change from 100 to 110 mg/L is only a 10% increase, equivalent to a change from 10 to 11 mg/L. A doubling of toxic strength in the higher range would be from 100 to 200 mg/L. If the preceding argument fails to

[1] For example, Robertson *et al.* (1984) found that precise tests to determine LC50 required concentrations to be equally spaced on a log scale, particularly for those concentrations yielding effects from 25 to 75%. They were conducting specific studies of the needs for an effective toxicity test, using insects for the trial tests.

convince, one might consider it as a question of the units of concentration. The series 1, 2, 4, 8 mg/L is *identical* to the series 1000, 2000, 4000, 8000 µg/L, even though the absolute changes appear to be vastly different at first glance.

Accordingly, use of the logarithm of concentration reflects a biological phenomenon, fits toxicological exposure, and is *not* a transformation adopted primarily for statistical convenience.

Sometimes, investigators who normally use a geometric series of concentrations, but wish to have more detailed information in a particular limited range within the total series, have been known to abandon the geometric principle and use an arithmetic series. For example, they might cover a range of particular interest from 30 to 60 with the concentrations 30, 40, 50, and 60. The same principle applies, that this should have been a geometric series. The intervals are uneven, that from 30 to 40 being a 33% increase, while that from 50 to 60 is only a 20% increase. Usually, it would have been more appropriate to have divided the entire range of the test, including the section of most interest, into finer equal intervals.

There are a few exceptions to logarithmic transformation of dose. One would be pH which is already logarithmic. Another would be temperature, which is a special case, and for biotic interpretations has nothing equivalent to the zero of the concentration scale of a toxicant.

2.3.1 Maintaining Logs

Although Canadian investigators seem to easily adopt a geometric series of concentrations for exposure, they are often reluctant and sometimes actively hostile to the idea of continuing with logarithms for statistical analysis. The reason is not clear, but might have to do with the increased arithmetic complexity, and/or lack of familiarity with logarithms (see Section 2.3.5 on familiarization). It is a common mistake to conduct statistical analyses using arithmetic values of concentration. If the results seem satisfactory, investigators see "no need" to use logarithms, and they proceed with arithmetic values. As described in the following text, that approach is backwards because it abandons the initial scale without showing cause. The proper procedure is to start the analysis with log concentrations. If the requirements of the procedure are satisfied, and results are satisfactory, the scale of concentrations is retained.

It is simply good science for an investigator, having adopted the geometric/logarithmic scale for testing, and thereby having rejected an arithmetic model, to retain that scale throughout the investigation and analysis, unless it is proven wrong. Adopting the scale is akin to adopting a hypothesis -- one sticks with it until it is demonstrated to be incorrect, in which case one looks for a better hypothesis (or in this case, a better scale of concentrations). This is not primarily a matter of toxicology, or of statistics, but of science and the scientific method. The geometric series of concentrations has a fundamental "truth" in its use, (see opening paragraphs of Section 2.3), and statistical analysis should retain that fundamentality if the effects are to be interpreted without distortion. Even statisticians might occasionally overlook this basic reason for maintaining logarithms of concentration in analyses of results, and biologists or toxicologists should be prepared to support the concept. Published statements from statisticians acknowledge that the model for analysis should follow "underlying scientific reason", which is interpreted here as the rationale for adopting a geometric series for exposure (e.g., Collett, 1991, p. 94) [2].

Dropping the geometric scale part way through an investigation is common, unfortunately. It might help if an investigator asked him/herself: why were the initial exposures laid out in logarithmic series? *Whatever reason that was*, remains valid throughout statistical analyses, until proven to be incorrect.

If formal statistical testing showed that the model did not fit the data satisfactorily, then it is possible

---

[2] Collett writes of whether or not to use logarithms for probit or logit analysis. "In the *absence of any underlying scientific reason* for using transformed values of an explanatory variable in the model, the choice between alternative models will rest on statistical grounds alone and the model which best fits the available data would be adopted." [Italics added] Although at first glance, this statement might seem to give primacy to statistical considerations, that would only be so if there were no scientific (biological) reason to adopt a particular model.

that the logarithmic scale might be unsuitable.[3] The fit could be tested for an arithmetic or some alternative transformation of concentration, to seek a suitable fit (Section 2.9). If the alternative scale of concentration was shown to be superior, then the test should really have used that series for the exposure concentrations. However, at the risk of repetition, it is the logarithmic values of concentration that represent the default, and arithmetic values *must not* be taken as the default and tested first for suitability.

Retaining logarithms of concentration means that all subsequent mathematical manipulations of the data should be logarithmic (Sections 2.3.2 to 2.3.4).

2.3.2 Logs in Computer Programs

Investigators using a computer program designed for toxicological testing *must* satisfy themselves that log concentration is used in calculations.

Most or all of the available computer programs assume that exposures are done using a logarithmic series of concentrations, as seen in their example sets of data. Most programs automatically retain logarithms for probit regression, but not necessarily for other types of data analyses. The programs differ, and it might not be easy to discern what scale of concentration is used. Unaccountably, one older commercial program apparently had arithmetic concentration as the default; a trial of TOXSTAT 3.5 indicated that this was so, even for probit regression. There was an option to choose log concentration, but the investigator had to immediately enter another command, "RUN", otherwise the instruction was ignored. If all else fails with a commercial program, the investigator should enter the concentrations as logarithms. Any spreadsheet will provide the logarithms, but many of the commercial programs for toxicity analysis require that each item from the set of data is entered (tediously) into a particular segment of the program.

Surprisingly, the program ICPIN fails to use logarithmic concentrations for "linear interpolation" to estimate the ICp in sublethal quantitative tests (Section 6.4; Norberg-King, 1993). The procedures in the program were initially set up by personnel from the USEPA, and are now incorporated into commonly used commercial programs.

Computer programs for the newer approach of nonlinear regression are general-purpose, and not designed for toxicology. These programs have no automatic provision to use log concentration, and some authors using nonlinear regression have failed to change the arithmetic values (Section 6.5.7). Although nonlinear regression can fit almost any shape of curve, there will probably be a penalty of requiring more parameters, with loss of power for the fit (see Section 6.5.5). Usually, a model with logarithmic concentration and/or time will have a less complex relationship. The model can fit a simpler curve or straight-line relationship, with fewer parameters to estimate, and fewer degrees of freedom lost, resulting in a more powerful analysis. In addition, arithmetic effect-curves and graphs could be misleading (Section 5.3).

2.3.3 Logs in Further Calculations

Once calculated, endpoints with their confidence limits are often converted to arithmetic values for ease of comprehension. However, before any further mathematical manipulations are made with these arithmetic versions of the endpoints, confidence limits, or associated variables, they must be changed back to logarithms. (This is the "dose metameter" of Finney, 1971.)

A common mistake is to calculate an average of the arithmetic values of two or more EC50s, ICps, or other endpoint. One must remember that the C in EC50 and ICp represents *concentration*. The test endpoint should be thought of as a logarithm, sometimes temporarily transformed into an arithmetic value. Proper procedure is to average the logarithmic values of the endpoints, then if desired, take the anti-log of the result (a geometric mean) [4].

---

[3] The investigator must bear in mind that a failure of the model to fit the data might not be attributable to the transformation of the independent variable (the concentration), but instead, to one or more of the following causes. (a) Transformation of the effect is necessary. (b) The tolerance distribution (see glossary) is not normal. (c) The tolerance distribution has different scales (variances) at different concentrations. (d) The choice of model is incorrect for the data.

[4] An excellent example of the extensive averaging of endpoints, and other manipulations of data using the proper logarithmic methods, is provided by the USEPA in the document for developing water quality criteria (Stephan *et al.,* 1985).

In assessing reports of toxicity studies, it continues to be necessary to check for this type of error.

2.3.4 Does it Matter?

Some investigators protest the use of logarithms, because results are similar with arithmetic values of concentration. Although reasonably true for some "good" sets of data, there can be appreciable differences for the irregularities often seen in environmental toxicology.

If logarithms give a truer estimate for irregular data, the principle is established that the correct procedure is to use logs for all sets of data. The question is not "does it make much difference?", "but which is correct?" Canadian investigators should use the correct way.

Two examples that give credence to the use of logarithms are provided in Appendix D. The first example is simply a comparison of arithmetic and geometric averages for some sets of endpoint concentrations. For consistent ("good") data, there was little difference between arithmetic and geometric means. However, the two types of means diverged more and more as the sets of data became more irregular. In an extreme case the arithmetic mean was 5.4 times higher, and not representative of most values in the set.

In the second example, EC50s were calculated for the four data-sets A to D in Table 2. Running the probit regression with arithmetic concentrations yielded EC50s that averaged 1.2 times the proper values. The confidence limits were also raised to generally higher spans. Another example with a similar magnitude of error is shown for improper calculation of warning limits of reference toxicants (Section 2.8).

Elaborate mathematics might minimize this type of error, but *can never eliminate the fundamental flaw in approach*.

Similar errors could apply to subsequent manipulations which failed to use the logarithmic version of endpoints (e.g., means, trends with time, comparisons of potency, ANOVA, etc.). This might lead to false classification into pass/fail categories, or actions taken on differences which were not real. An incorrect toxicity curve could lead to a false conclusion of a threshold of effect, as shown by an example in Section 5.2.

2.3.5 Familiarization and Techniques

Because electronic calculators and computers have been easily available for several decades, modern investigators are often unfamiliar with logarithms and their structure. It can be beneficial to spend a little time studying the nature of logarithms. Exploration using a hand calculator with a logarithmic/antilog key will provide rapid insight. Of particular interest would be study of arithmetic manipulations versus their logarithmic equivalents:

- multiplication/division equals addition/subtraction of logarithms
- square roots and other roots can be done by division of logarithms.

The glossary gives some additional insight with examples of the format of logarithms.

Investigators who are uneasy with use of a logarithmic scale should consider that hydrogen-ion concentration in water is customarily described as pH, a logarithm, and seems easily accepted by most people.

There are some difficulties and inconveniences in using logarithms, but they can be circumvented. There can be a problem of entering the concentration for control data, because zero concentration does not have a logarithm. This could raise difficulties in estimating ICp by the currently available program ICPIN, which requires a concentration to be entered for the control [5]. The remedy is to enter the logarithm for some very low concentration in relation to the tested concentration (say 0.001 mg/L). A program such as ICPIN does not actually use that value, but identifies the control effect by its position in the table of data, so no harm is done to analytical procedures.

---

[5] The ICPIN program fails to convert concentrations to logarithms, so Canadian investigators must enter the logarithms of all concentrations. That includes entering the logarithm of a very low concentration for the control in the second row of the table of data.

Concentrations less than unity (1.0) have negative logarithms and tend to be disorienting. For toxicologists, the best remedy is to change the units of concentration. If test concentrations run down to, say, 0.1 mg/kg, the use of micrograms per kilogram will change the values to 100 and up, yielding positive logarithms. If the troublesome values were percent concentrations, the scale could be changed to parts per thousand, or parts per ten thousand. After the calculations, the results could be converted to arithmetic values for ease of comprehension, and the units could be modified back to whatever was desired. Modern computer programs have no problem handling negative logarithms, but there have been older or "local" computer programs that could not. Therefore, a cautious stance would be to enter only positive logarithms into computer programs.

2.3.6 Log Time

In environmental toxicology, time is usually part of the dose as well as part of the effect or response, and must also be considered in terms of log time. The geometric/logarithmic nature of time is not so self-evident, but the rationale parallels that for concentrations. It is not the absolute change in time that governs a change in effect, but the proportional increase in time. In a toxicity test, an increase in exposure from one hour to two hours would represent a doubling of exposure, perhaps with a major change in effect. An increase of one hour from 96 to 97 hours would represent a trivial increase, probably not detectable as any change in effect. Accordingly, logarithms of time should be considered during test design and used in any analysis involving times.

There is some recognition of this, in that investigators are likely to make frequent inspections of a test at the beginning, then gradually lengthen the intervals of observation. This is a tacit acknowledgment that an hour at the beginning of a test is more important than an hour at the end of a one-week exposure. Psychologists note that humans' perception of passed time is logarithmic, to some extent (Cohen, 1964). An early aquatic toxicologist (Wilber, 1962) described the situation as follows.

> *"Biological Time*
> *When long-term studies are made in which toxicants are used in sublethal concentrations it is important to recognize that biological time is a logarithmic phenomenon [Du Nouy, 1936]. This fact has been called to mind by others [Gaddum, 1953]. It may partly explain why dose-response curves in which time is an element are of a logarithmic nature.*
>
> *The logarithmic character of biological time must be kept in mind when one interprets long-term experiments with water toxicants. It is evident that the biological value and significance of a given time interval will not be the same at the beginning of a chronic exposure as it will be near the termination. Such a consideration might well be important in modifying one's conclusions."* [6]

The use of time in toxicity analyses is primarily in toxicity curves (Section 5), but for quantal effects there are advantages in estimating times to 50% effect (ET50, Section 5.1).

2.3.7 Logarithm of Effect?

The independent variable in toxicity tests can sometimes have a logarithmic nature, and should be analyzed as such. This could arise when measuring quantitative effects. For example, when calculating an ICp for weight changes in organisms, the changes are assessed as proportions of the weight of the organisms. The ICp itself is calculated as a designated percent impairment, i.e., a proportional reduction from the control. In other words, the ICp deals in ratios, so the intervals are geometric or logarithmic.

Most of the arguments presented in this section, about using logarithms as the default for the independent variable (concentration), would also seem to apply to quantitative dependent variables which are proportional by their nature (e.g., weight). However, the concept is commonly applied in only one situation, i.e., in transformation of effects data to achieve conformity with requirements of

---

[6] Du Nouy, cited in this quotation from Wilber, wrote a book on biological time, and Gaddum was one of the early giants in pharmacological toxicology. (See Reference list.)

normality and homogeneity of variance (see Section 2.9).

Aside from that kind of transformation when needed, day-to-day work in environmental toxicology shows little recognition of the concept of a proportional scale of effects. Perhaps the topic will emerge as new methods are developed in the future. One of the most advanced toxicological statisticians (Slob, 2002) has adopted this approach in modelling quantitative data. Slob (2002) describes his assumptions in nonlinear regression: "As a default, it is assumed that the measurements are lognormally distributed. Consequently, the dose-response model is fitted on the log-scale, i.e., both the model and the data are log-transformed. ... Therefore, the group means are not arithmetic but geometric means ..."

2.4 Randomization

Randomization is somewhat analogous to insurance, in that it is a precaution against disturbances that may or may not occur, and that may or may not be serious if they do occur. Cochran and Cox (1957)

Key Guidance

- *Statistical tests assume that all ancillary variables in a toxicity test are random. Randomization should, therefore, prevail in all aspects of design and procedures. This includes randomization of containers for different concentrations, position of containers in an array, and placing organisms into containers.*
- *Another possibility of bias can be removed if the observer does not know the identity of the test containers.*
- *Practical methods of randomization are shown in Appendix E.*

In toxicity tests as in other experimental work, randomization is of critical importance for statistical inference. It makes the test assumptions valid by destroying any potential correlation among the experimental units. The independence of observations allows unbiased estimates of treatment effects and variances. Davis *et al.* (1998) concluded that "nonrandom allocation of organisms can produce significant bias in estimates of lethal concentration". Any reasonable attempt at randomization removed the bias, but the least variance for the result was obtained by completely random allocation.

Randomization should prevail in all aspects of the design and procedures for a toxicity test. Any statistical test assumes that all variables contributing to the data are random, except the variable being investigated, which in this case would be the toxic agent(s). If one of the ancillary variables is not deliberately randomized, there is automatically a question about the validity of statistical treatment. By failing to randomize an item, the investigator is assuming that the item will not bias the results or invalidate the statistical tests, which could be true. However, if it did cause bias or invalidation, there is usually no way of ascertaining that after the test. The only way to avoid the uncertainty is to randomize all the possible contributing factors, aside from the levels of concentration and exposure time that are selected to be part of the "dose".

If an Environment Canada test method has a "must" requirement for randomization, a deviation from that procedure must be reported, and might invalidate the test. For regulatory tests which might be used in legal proceedings, suitable randomization removes one avenue for criticism of the test (and the investigator) by any outside organization that wished to cast doubt on the results.

Elements that should be randomized include the following list.

- *Randomization of containers used for concentrations* is seldom done, but should be. If a container had been used in a previous test, it is possible that a toxicant might carry over, even with cleaning, and influence the new effect observed in that container. Even new containers could conceivably have some occasional flaw or component that would affect a test in an irregular way.

- *Random placement of containers within the room, incubator, etc.* is specified in most test methods published by Environment Canada. There could

be differences in ancillary conditions according to the particular area. Sometimes there is resistance to this procedure because an irregular array of concentrations and replicates could cause errors in recording data (see below).

- *Random placement of test organisms into containers can be important.* Often it is not done because it can be tedious, and sometimes difficult to keep track of how many organisms have been put into a given container. Formal randomization can be done, or even a system like dealing out cards can be satisfactory.

- *"Blind" tests, in which the observer does not know the treatments* means that the containers must be identified by some code, rather than being labelled with their concentration. A blind test represents a high degree of care in avoiding observer bias, and would contribute to an unassailable test result.

The worst situation would be a *systematic bias* because of failure to randomize. For example, if the test organisms captured from the colony were used to fill the concentrations in the order that they were captured, the most-easily caught organisms might go to the lowest concentrations; possibly they might be weaker and more sensitive to toxicants, resulting in an exaggerated effect at low concentrations. Similarly, if the test containers were lined up in order of concentration, results might be biased by a gradient of temperature, light, or disturbance which was present in the laboratory. For example, proximity to a heater in an incubator could influence test temperatures, and hence toxicity. Algal tests can be particularly susceptible to variation because growth falls off steeply at reduced levels of light, which might be found at the edges or corners of an array being tested. Even with excellent randomization, unrecognized outside conditions might influence toxicity in certain test containers, but *that would merely increase the general variation of the test result, without a systematic bias*.

A complicated procedure for randomization could contribute to a definite risk of operator error in assigning exposures or recording data. Certainly it could result in extra time and trouble. Even statistical experts from the OECD (2004) recognize that in "some circumstances it may be difficult, or expensive, to randomize at every stage in an experiment." If some randomization must be omitted, they recommend a separate examination of the potential effect on the test results. For these reasons, some tests in Canada are probably deficient in randomization, and investigators should realize that their results could be biased. If randomization procedures are deliberately compromised for good reason, it should be done in such a way that *only the total variation of the test is likely to be affected*, by attempting to minimize the possibility of a systematic, concentration-related bias. Usually, the only way to be completely sure of avoiding that systematic bias is complete randomization in each step of the test.

Some helpful procedures are given in Appendix E, for putting organisms into containers, and containers into positions. Most statistical texts offer advice and methods (e.g., Fleiss, 1981).

2.5 Replication and Numbers of Organisms

Key Guidance

- *Replication allows the variation within each concentration to be evaluated, which, in turn, can be used to decide on significant differences among concentrations.*
- *In a given test, a replicate must be an independent test chamber containing one or more organisms, with no connection to another chamber through the test medium.*
- *A treatment includes all the replicates at a given concentration, and all the organisms within each of those replicates.*
- *Correct understanding and use of terminology is important, otherwise statistical tests could be used in ways that are invalid.*
- *The number of organisms per concentration or replicate is an important factor in design. Practical limitations in the laboratory can prevent the use of enough organisms to achieve statistical ideals. Replicates could be an advantageous way of providing suitable*

conditions for test organisms, or security in case of accident with one container.

- *When a quantal test is analyzed by probit regression, any replicates would be pooled. Replicates are beneficial, however, when using more sophisticated statistical tools.*
- *If regression is used for a point estimate with quantitative data, replication allows testing the goodness of fit, and the divergence of the model from the data. Point estimates by smoothing and interpolation can be done without replication, but the commonly used ICPIN program requires two replicates, and preferably five, to assess significance. Replication is an essential part of hypothesis testing.*
- *When samples are collected for testing, field replicates (true replicates) are separate samples of sediment, water, etc. from the same time and general location. They are excellent replicates in a toxicity test, to incorporate the variation in the substrate being assessed. Subsamples of one sample ("laboratory replicates") assess variation in laboratory technique and homogeneity of the sample, but do not provide any information for distinguishing field locations.*

2.5.1 Terminology

Using the correct terminology in toxicity tests can be important. Incorrect usage might cause a statistical test to be applied incorrectly, resulting in conclusions that are invalid.

A *replicate* in a toxicity test is a single test chamber containing one or more organisms, and it is one of two or more such chambers exposed to the same *treatment*, i.e., exposed to the same concentration of test material (or the control condition) [7]. Thus replicates are repetitions of the *experimental unit*, the smallest independent element in a toxicity test, to which a treatment is applied. These terms are further explained in the following text, and in the Glossary.

There could be only one organism in a test chamber, and that would still be a replicate and also an experimental unit. An example is provided by Environment Canada's test of survival and reproduction of *Ceriodaphnia* (EC, 1992a). Each of the 10 parent organisms in a treatment is a replicate and also an experimental unit because it is in a separate test vessel. The test counts the number of young produced by each organism.

However, since individuals show differing sensitivity, a single organism in a replicate means that the replicates are just as variable as the organisms (hence the large number of ten replicates in the test with *Ceriodaphnia*). Several organisms per vessel are normally used to improve the precision. The organisms in a chamber are *sampling units* providing data that contribute to the result for the replicate.

A replicate must be independent. The separate chambers which are replicates must have no connection between them through the test water, sediment, or soil. Thus, if several permeable chambers in an aquatic test were exposed by suspending them in one tank of test solution, the chambers would not be replicates. Similarly, test material which has contacted one replicate chamber must not be transferred into contact with another chamber. Nor can there be any transfer of organisms between chambers, once the test has started. Failure to meet these requirements would invalidate a statistical analysis based on replication.

There is some variation in terminology used in environmental toxicology. The term "replicate treatment" has been used in some methods documents published by Environment Canada to mean the same as expressed here by *replicate*. "Replicate treatment" is confusing since it combines two levels of a hierarchy (see following text), and *replicate* is recommended instead. Statisticians sometimes use the term *replication* as a noun representing one test chamber (Snedecor and Cochran, 1980), thus they could speak of several

[7] The examples of *treatment* in this section are all associated with concentrations, but that need not be so. A sample of sediment collected from the field would also be a treatment, when it was tested.

"replications" for a given concentration, meaning several chambers given the same treatment. This word seems better used as a general noun representing the practice of creating replicates.

Investigators should be alert for any mistakes in the instructions for computer software, which have occasionally referred to individual organisms in a chamber as "replicates". They are not, and pretending that they are would be one form of *pseudoreplication*. Such organisms are *sampling units* or *subsamples* contributing to one replicate. In common parlance, information from an organism might be called a "measurement" or "observation", for example, "the ten measurements in the first replicate were ...". Some additional comments on this mistake follow under "Hypothesis testing" (see also, Section 7.2.1).

The appropriate use of terms can be illustrated by an example of an ordinary sublethal aquatic toxicity test.

4 test concentrations and a control	= 5 *treatments*
2 isolated test chambers for each concentration	= 2 *replicates* per treatment
6 fish in each chamber	= 6 *sampling units* per replicate
Overall: 5 treatments with 2 replicates	= 10 *experimental units*

Accordingly, an experiment can have three levels of variation in the measurements:

- among individual organisms in a container *(the sampling units)*;
- between individual containers at the same concentration *(replicates)*; and
- among concentrations *(treatments)*.

Clearly an investigator must understand the differences, particularly when doing an ANOVA.

2.5.2 Replication in Various Kinds of Tests

Replication of test chambers can be a powerful way of improving the output from some toxicity tests. It allows an assessment of the variation or "noise" within each concentration, and it allows a statistical test for lack of fit. Hurlbert (1984) is a highly recommended paper on replication.

Replicates in quantal tests. Replicates at each concentration are usually not necessary, because all the results for each concentration are combined, before estimating the LC50 or EC50 by the classical methods such as probit regression commonly used today. Replicates are sometimes convenient or useful, however, for handling and providing suitable conditions for test organisms. For example, dividing the total number of organisms at a given concentration into several replicates would be a suitable way of providing the required volume of test material in a container of convenient size.

Also, there could be a real benefit of "insurance" in a test, in case of accidental damage to one chamber, loss, or disease. If one replicate suffered such misfortune, the others would usually be suitable for use in analysis of results. As an example of this, Environment Canada requires three replicates in the sublethal/lethal test with early life stages of rainbow trout (EC, 1998a). The test would not seem to require replicates, because it emphasizes the quantal endpoints EC50 and EC25 for non-viability and development. The reason is an appreciable risk of damage or disease in working with the delicate eggs and young stages of trout, and the replicates increase the likelihood of getting suitable data at each concentration [8].

Replication is beneficial if more sophisticated statistical programs are applied to quantal tests; such programs might become available in the future for general use.

Numbers of organisms in non-replicated quantal tests. Increasing the number of test organisms can improve precision in a test, thus achieving narrower confidence limits on the endpoint. In quantal tests, the ratio between confidence limit and EC50 could

[8] Replicates are required for other tests with quantal endpoints such as the test using fathead minnows (EC, 1992b). Because the tests are dual-nature, with lethal and sublethal components, replicates are required for the latter.

be cut in half by using 30 organisms per test chamber instead of 10 organisms (Hodson *et al.*, 1977). A similar improvement had been quantified by Jensen (1972), who found major decreases in variance of the LC50 as the number of test organisms increased from 1 to 10 per treatment. There was a further decrease of 29% in standard error when organisms increased from 10 to 20, a decrease of 13% as test organisms increased from 20 to 30, then only an 8% improvement for a change from 30 to 40 organisms. Improvements were small for more than 30 test organisms per treatment, in these tests of lethal temperatures. Of course, the exact results for comparisons like these will depend on the spacing of concentrations around the LC50.

Statisticians urge an increase in numbers to improve precision, but other factors also affect the choice of number of organisms, e.g., economy, size of tanks, available volume of test sample, and animal rights legislation. In tests with fish, there is a current trend towards using fewer per chamber, partly to reduce the destruction of living organisms. Douglas *et al.* (1986) indicated little loss in precision by a 44% reduction in number of test organisms, achieved by using seven animals at each of four concentrations, instead of ten at each of five concentrations. However, the reduced number of concentrations leads to danger of missing the important effect range (Section 2.2), and beyond doubt, there is some loss of precision if the number of test organisms is decreased to less than 10 per treatment, as indicated in the preceding paragraph.

Point estimates by regression. In the current trend towards regression for estimating endpoints of quantitative sublethal effect (Section 6.5), it can be beneficial to use additional concentrations that are more closely spaced (Moore, 1996; Section 6.2.3). Accordingly, there is some pressure to use resources for more concentrations, rather than more replicates. Indeed, conventional regression analysis requires, strictly speaking, only one measurement at each concentration. At its simplest, *regression analysis* describes the linear relationship between an observation such as size, and an independent continuous variable such as the logarithm of concentration. After the relationship is defined mathematically, it is used to calculate the endpoint. Confidence limits on the endpoint can be obtained with or without replicates.

Nevertheless there are major reasons for having substantial levels of replication. Environment Canada has recommended 3 to 10 replicates or more, in recently published test methods which require regression techniques (EC, 2004a–c and Appendix O). The primary reason is that replicates are essential for assessing the fit of a regression [9]. Without replicates, there is no way of distinguishing the error derived from the scatter of observations at the same concentration (call it *pure error*), from a real divergence of the data from the pattern of the model (call it *lack of fit error*) [10].

Smoothing and Interpolation. If the ICPIN method (Section 6.4) is to be used for estimating the ICp, then at least two replicates are required to calculate the confidence limits. Each replicate contributes one measurement, for example, the average weight of organisms in that replicate. Five or more measurements (replicates) per concentration would reduce the width of the confidence interval.

Hypothesis testing. Replication is essential for analyzing results by *hypothesis testing*, once a favoured approach (Section 7). Larger numbers of replicates are beneficial for analysis of variance, allowing greater certainty in distinguishing the NOEC from the LOEC. If an investigator intends to carry out hypothesis testing as well as making a point estimate, more replicates could be added in the

---

[9] In some earlier methods published by Environment Canada, testing the goodness of fit was not a firm requirement. It would be a decision of the investigator, who might wish to document that the regression model was a suitable fit.

[10] An important benefit of replication is to distinguish between two categories of variation in a given test. The *pure error* would be the apparently random scatter, caused by the different sensitivities of individual organisms at the same concentration. The other category would be *lack of fit error*, a consistent pattern of divergence from the chosen regression model. Replication is necessary to separate those two categories of variation.

An example of a *lack of fit error* would be when a straight-line pattern had been adopted as the supposed pattern of concentration-effect, but the data represented a convex curve that increasingly diverged from linearity at higher concentrations.

design of the test. Environment Canada requires at least four replicates if NOEC/LOEC is to be estimated in the sublethal test with young stages of rainbow trout (EC, 1998a). Those four replicates might be essential if parametric analysis were invalid and nonparametric methods were required.

Hypothesis testing has a particular danger from *pseudoreplication* (see Section 2.5.1). It is not difficult to imagine the gross errors in conclusions that could arise if the organisms in a container were mistakenly entered as replicates into an ANOVA. If there were, for example, 10 worms in each container, the statistical test would mistakenly treat this as a powerful experiment indeed. Random differences might emerge from the analysis as if they were significant ("real").

2.5.3 Inter-relationships with Field Sampling

When samples collected in the field are tested in the laboratory, there are some relationships between the test procedures, and interpretation of results back to the field. This would be especially relevant when samples of sediment or soil ("substrate") were brought into the laboratory, but might sometimes apply to water samples [11].

In particular, there is a very important difference between test replicates based on separate samples of the test material, and replicates based on subsampling of a single sample. Samples that were *field replicates* would be separate samples of soil, sediment, etc., collected in the field by identical methods and at the same sampling station. The purpose would be to allow the investigator to evaluate the variation in quality (or qualities) of the sampled substrate at that station. This kind of sample is sometimes also called a *true replicate* or a *replicate sample*. The field replicates must be stored in separate containers and, as often recommended for tests of soils or sediments, each can be used to set up a replicate in each treatment of a toxicity test. The procedure would incorporate into the toxicity test, (a) the variation in the sediment or soil at a given station (and variation in sampling it), combined with (b) any variation created by conditions or procedures in the laboratory.

Subsamples could be created in the laboratory by dividing a single sample of substrate. These are also called *laboratory replicates*, but "subsample" conveys their nature. If such subsamples were used as replicates in a toxicity test, the results would assess the homogeneity of each sample and the

---

[11] This document does not offer guidance on field work, but some further comment on sampling is relevant to setting up and interpreting toxicity tests. Sometimes there can be considerable uncertainty in deciding what represents a replicate in field sampling, for example in evaluating the sediment of a bay. The general principle is that replicate samples should adequately cover the area considered to be uniform, which the investigator wishes to characterize. If the entire bay is to be characterized as a single unit, then samples collected at a number of points around the bay would be *field replicates* (true replicates or replicate samples). Under these circumstances, if a number of samples were collected at one point in a bay, they would not really be field replicates representing the variation in the entire bay, but subsamples at a particular location within the bay.

On the other hand, if the investigator wished to assess the effects of pollution in different parts of a bay, there would be a rather different sampling strategy and a different outlook on replication. There might be a set of samples at one sampling station at the head of the bay, near a point-source of pollution. Another set might be at a station in the northerly outer part of the bay, to assess the effect of dilution as the effluent was carried outwards by a current circulating around the bay. A third set of samples might be taken at a station in the southerly outer part of the bay where incoming new water was expected to be clean. If several samples of sediment were taken at each place, the samples at a station would be replicates. The purpose would be to determine if the three stations had differences in pollutional status, differences which were significant against the background of variation measured by the replication at each station.

Clearly, valid conclusions from the toxicity study would require the field sampling to be based on good understanding of physical factors in the habitat of interest. For example, in the bay used above as an example, there might be different water movements in deep and shallow water. In any plan for sampling sediments, the two depths would need to be treated as different areas, in addition to the areas represented by horizontal location about the bay.

Such distinctions in replication are relevant to Canadian programs of Environmental Effects Monitoring, in which the field surveys are tied into toxicity tests in the laboratory.

variation due to testing procedure, and that might be an appropriate feature of the design. However, subsamples *would show nothing about variation of the substrate in the field* (e.g., sediment in a lake, see Section 3.1.3). Depending on the purpose of the investigation, effort devoted to preparation and analysis of subsamples might be more profitably spent in obtaining field replicates. Accordingly, laboratory subsamples are not usually recommended here, unless they facilitate handling of the test organisms (fewer per chamber), make it easier to set up the tests (e.g., smaller containers), satisfy a need to assess homogeneity of the sample and variation in laboratory technique, or are required by a particular test method.

Environment Canada (1994) is an excellent guidance document on replication of sediments.

2.6 Weighting

Key Guidance

- *"Weighting" certain observations gives them more influence on the results of subsequent calculations.*
- *A value is weighted because it is thought to be more valuable for one of several reasons: (a) it is close to the endpoint of interest; (b) it represents many organisms or measurements; or (c) it represents measurements with a small variation.*

The Glossary indicates that *weighting* of a certain value within a series, means that an arithmetic manipulation is applied to that value, to change its influence on whatever calculation is being carried out with the series. Common reasons for weighting would be unequal numbers of measurements in the groups of a series, or unequal variances within a series. The following paragraphs expand on the uses of weighting.

One example of weighting is provided in Section 4.2.3, on using a hand-drawn graph to estimate an EC50. The advice is that in "fitting a probit line by eye, ... the points should be mentally weighted. Those closest to 50% effect should be given the most weight ...". From a practical point of view, the central points are weighted because that portion of the data-set is closest to the endpoint of interest, and most likely to estimate it accurately. Informal weighting of this kind is subjective, to say the least, but is better than ignoring the relative value of points on the graph.

The concept of formal weighting might be introduced by a simplistic example of mathematically fitting a line. If the values thought to be most valuable were entered twice into the data-set, they would have more influence on the fit, i.e., they would carry more weight. (It need scarcely be said that this is in no way an allowable procedure, and is merely used to convey the idea of weighting.) Formal weighting is often quite sophisticated, as in probit regression (Section 4.5), where it is based on expected probit and has continuously variable magnitude.

A common use of weighting is to compensate for the number of measurements which contribute to a given value in a series. If each value in the series was the mean of measurements on a sample of organisms, a particular mean might be weighted because it was based on a large sample of organisms. This would be done before the analysis was carried out. The adjusting factor might be as simple as the number of organisms.

The mean of a group of observations might also be weighted because it was derived from observations which showed a small variance, therefore making the mean appear to be an especially valuable estimate in a series. If the group of observations itself was being used in analysis, weighting could be applied directly. This kind of weighting is essential when fitting a model to data in which some groups are more variable than others. The model will almost certainly require equal variances. The observations can be weighted in a clever mathematical way so that the assumption of equality of variances is restored; usually the computer program of a model handles this step.

Specific reference to weighting follows: Sections 4.2.2, 4.2.3 and 4.5.1 to 4.5.3 (various aspects of probit regression); Section 4.5.6 (Spearman-Kärber estimates); Section 4.7 (nonlinear models for quantal data and smoothing for kernel methods);

Section 6.4 (ICp by smoothing and interpolation); Section 6.5.4 (inverse of variance, for nonlinear regression); and Section 8.2.3 (unequal numbers of replicates in dual-effect tests).

2.7 Controls

Controls in a toxicity test represent a treatment that duplicates all the physical, chemical, and biotic factors that might affect the results of the test, except for the specific condition being studied. None of the material being tested for toxicity is added to the control. The control is used as a baseline of experimental effects resulting from conditions such as the quality of dilution water or health and handling of organisms. *Control* is synonymous with *negative control*.

A particular test method might require replication of each test concentration. In that case, there must be the same replication of each type of control. Some test methods specify different numbers of replicates for toxicant and control (Appendix O, Table O.1).

Key Guidance

- *A control must be identical to test concentrations in every way except that it has none of the material being tested for toxicity (i.e., a zero treatment or the null treatment). It provides a baseline for observed effects.*
- *If a solvent is used to get a test chemical into solution, then a solvent control must be used, containing solvent at the highest concentration used elsewhere in the test. This solvent control must not cause greater effect than the standard control.*
- *If salinities are unadjusted in a marine test, there should be salinity controls matching the salt content of the various treatments. If test waters are adjusted to a favourable salinity (30‰), the control is also at that salinity. Extra salinity controls are needed if test waters are adjusted by methods (dry salt or brine) that differ from the method for the control.*
- *Tests with sediments and soils use controls that follow the same principles as those for other tests. Comparison of test results is normally with a reference sediment/soil, collected in the field and thought to be clean. A control sediment/soil is also run to judge the overall quality of the test; it is manufactured or taken from a different and clean site.*

2.7.1 Ordinary Controls

Controls must be set up in exactly the same way as the test concentrations. Selection of organisms must be done at the same time and in the same way. The containers must be of the same type, and the dilution water, control sediment, or other substrate must be uniform throughout all the containers. Controls must be randomly arranged among the other containers. Only in this way can there be an unbiased assessment of whether there is an effect caused by something other than the test material or treatment.

The controls serve as a baseline, but the methods documents published by Environment Canada insist that the baseline must indicate satisfactory conditions and procedures. The specific requirements for control performance vary with the type of test, but some examples can be given. In the test for growth and survival of polychaete worms, there must be at least 90% average survival in the controls (EC, 2001a). For larval growth of fathead minnows, no more than 20% of the control larvae can be moribund or show atypical swimming at the end of the test, and they must also attain an average dry weight of 250 µg (EC, 1992b). When testing embryos of salmonids (EC, 1998a), the average percentage of nonviable control eggs must be no greater than 30%. With duckweed, the number of fronds must have increased by at least eight-fold in the controls (EC, 1999b).

2.7.2 Solvent Controls

Sometimes a chemical that is being tested for toxicity is only sparingly soluble in water. A solvent could assist in getting the higher concentrations needed for a strong effect in the test.

Usually, this would be relevant for aquatic toxicity tests (fish, algae, etc.), or sediment tests in which a chemical was being "spiked" into the sediment. It might also be relevant for a toxicity test with soil, if the chemical was added to the soil by means of a solution of that chemical.

In EC toxicity tests, the preferred option is to use only dilution water as a carrier for the test chemical; the use of any other solvent "should be avoided unless it is absolutely necessary" (EC, 1997a, b; 2001a). If assistance is needed for a sparingly soluble test chemical, the first choice is a generator column (Billington *et al.*, 1988). Less desirable would be ultrasonic dispersion, and even less recommended are organic solvents, emulsifiers, or dispersants (EC 1997b; 1998a; 2001a). Sometimes the requirement is stricter, in that no solvent should be used in a test unless it is one that might be formulated with the test chemical for its normal commercial purposes (EC, 1992f; 1998a; 1999b).

For design purposes, an EC test that uses a solvent must have a *solvent control*, i.e., a test chamber (or replicate chambers) that is just like the standard control except that it contains the solvent at the highest concentration that is present anywhere else in the test. This is in addition to the usual control. Needless to say, the concentration of the solvent should be well below its toxic level, and sometimes additionally, must not exceed 0.1 mL/L (EC, 1992f; 1999b). If unknown, its toxicity should be tested in the usual way to determine its threshold for effect, before it is used in any other test (EC 1997a, b; 1999b).

Effects of the solvent control must not be greater than in the (standard) control. Some EC methods have that requirement for solvent controls, with no specific statistical procedures designated (EC, 1992f; 1998a; 1999b). In some methods (EC, 1998a; 1999b), if there is a solvent control, then it automatically becomes the overall control for assessing the effect of the toxicant. However, in the sediment tests with midge larvae, *H. azteca*, marine amphipods, and polychaete worms, the solvent control is only used in that way if its endpoint differs statistically from the standard control (EC, 1992e; 1997b; 2001a). It is not desirable to pool data from the solvent control with those from the ordinary control, and that prohibition is a "must not" in the test with young stages of trout (EC, 1998a), because "the control/dilution water lacks an influence [the solvent] that could act on organisms in the other concentrations". Although in the sediment test with polychaetes (EC, 2001a), the data from the solvent control are to be pooled with the ordinary control data if the two are not different by *t*-test, that pooling might well be omitted. The OECD (2004) does not favour such pooling, and points out that real differences between the two controls might not have been detected by the statistical test.

In any test method, organisms in the solvent control must meet the performance criteria for test validity which normally apply to the control.

2.7.3 Salinity Controls

A *salinity control* is a separate control chamber or set of chambers, intended to assess the effect of less-than-optimal salinities in a toxicity test with marine organisms. It also serves the purposes of a normal control. There could be a need for salinity controls in tests with any marine organism, whether in a medium of water or sediment.

Tests with unadjusted salinities. A salinity control would be desirable if the investigator wished to run a test without adjusting salinities. For example, it might be desired to assess the total impact of a freshwater effluent, when it was discharged into a marine location. In such a case, a separate set of salinity controls should be run, in addition to the control of control/dilution water at a favourable salinity (30‰). These extra containers should have salinities that duplicate or span the salinities in the test chambers. Environment Canada stipulates that the salinity controls should be made up by adding to the saline control/dilution water in a series of vessels, distilled or de-ionized water at the same concentrations as those of the liquid being tested for toxicity (EC, 1992f). The same procedure would be logical if one were testing the toxicity of a sediment (say, dredge spoils) which contained a liquid component that was essentially fresh water, intended for deposit in a marine location.

Obviously, the purpose of the salinity controls is to indicate any deleterious effects of low salinity

acting alone. These controls would not, however, indicate any worsened harmful effect caused by combined actions of the divergent salinity and the test material. To interpret results, the only option would be to credit to the test material, any toxicity *greater* than that found in the salinity controls.

Tests would "normally be carried out without adjustment of salinity" in the EC test for sticklebacks (EC, 1990b). Although a salinity control is not required, it would be beneficial. The method provides an option of adjusting salinity to 28‰, for tests with chemicals, effluents, leachates, and elutriates.

Conceivably there could be a situation in which the unfavourable salinity was too high. Perhaps an effluent could be very briny, and suspected of containing toxic materials. The principles for running and interpreting salinity controls would be the same as for a low-salinity situation.

Salinity-adjusted tests. The standard practice in marine toxicity tests is to adjust all the test concentrations to a single favourable salinity. That practice is the usual procedure with echinoids (EC, 1992f) and in sediment tests with amphipods (EC, 1992e). It is always the case under Environment Canada's program of Environmental Effects Monitoring, with the four tests using marine organisms (EC, 2001b). Environment Canada has adopted 30‰ as a standard favourable salinity for such adjustment.

In these tests there would be no "salinity control". There would be a normal control at salinity 30‰, prepared with the same material that was used to adjust the salinity of the test concentrations and/or the dilution water.

Special salinity controls. Another kind of salinity control might be required in marine tests within the Environmental Effects Monitoring program (EC, 2001b). This relates to the particular technique used in adjusting salinity.

Salinity of an effluent or a test concentration may be adjusted upwards by adding dry salts (reagent grade chemicals or a commercial mix), or hypersaline brine (HSB). Normally, the same material would be used to prepare all test concentrations and controls, in which case, no additional special controls would be needed. [The normal controls might be called "salt controls" if made up from dry salts, and "HSB controls" if made up from HSB (EC, 2001b)].

If, however, the dilution water used in the test concentrations differed in origin from the salt control(s) or HSB control(s) used in the tests, then a second control or set of controls would be set up, using the dilution water ("dilution-water controls") [12]. All of these treatments would be at a salinity of 30‰.

Statistical analyses of salinity controls. The principle for interpreting the controls is that each type of control, individually, must satisfy the performance requirements specified for the control, in the instructions for that particular toxicity test. For example, in a test within the Canadian EEM program, the "salt control" would have to satisfy the specified criteria, and so would the "dilution-water control", if both had been used. If any category of control did not meet the requirements, then the toxicity test would be considered invalid.

Such a failure is most likely to arise for a test in which salinity was not standardized for the various test concentrations. High concentrations of a freshwater effluent would mean low salinities in the test chamber. The corresponding salinity controls at those low salinities might fail to meet performance standards. The conclusion would be obvious: any effects seen in the higher test concentrations were likely due, in whole or in part, to the effect of low salinity. It would not be a valid test to determine the effects of the toxicant (effluent).

If all types of control satisfied the performance requirements, then the toxicity test would be valid.

---

[12] The dilution water might be uncontaminated seawater with salinity 30‰, whereas dry salts were used to adjust the test effluent to the same salinity. Or, the dilution water might be reconstituted from HSB and de-ionized water, while the effluent was adjusted with dry salts. Other combinations are possible. The principle is that the special controls are needed if there is any difference whatsoever in the manner of preparing the controls, compared to the test concentrations.

Use of the control results in analysis of the findings, would then follow whatever standard practice was specified in the particular test method.

2.7.4 Control and Reference Sediments and Soils

In EC toxicity tests for sediment or soil, standard procedure includes a *reference sediment/soil* with each test sample or set of samples from a given area (e.g., EC, 1997a). The reference sample is presumed to be clean, and results from the test sample(s) are compared with the reference results to see if there is any effect such as increased mortality or smaller size. Accordingly, the reference sample is used as a standard type of control for the test. The rationale is sensible, because this provides a site-specific evaluation of toxicity.

The tests also run a *control sediment/soil* with each batch of samples as a check on the general quality of the test and its organisms. A limit for acceptable performance is set in each Environment Canada method, for example, there cannot be more than 30% mortality in the sediment test with midges (EC, 1997a). The control sediment/soil is not normally used as the base of direct comparison for effects in the test samples. However, it would be used in that role if the reference sediment proved "unsuitable for comparison because of toxicity or atypical physicochemical characteristics" (EC, 1997a). The approach is a reasonable one.

The two types of controls are defined in the glossary but their characteristics might also be distinguished here. A *reference sediment* is collected in the field, within the general vicinity of the survey stations, at a site thought to be beyond the influence of the source of contamination under study. This reference sediment is presumed to be clean, and to have physical characteristics that closely match the samples under study. A *reference soil* is taken from a terrestrial location, but otherwise exactly parallels the characteristics and functions of a reference sediment. Because it is the control, it incorporates the *matrix effects* into the test. It can also be used as a diluent to prepare concentrations of the test soil.

A *control sediment* or *control soil* would not be collected in the same general vicinity as the survey samples. It could be taken from an uncontaminated site, or it could be made up from appropriate constituents. The objective is to get a clean sediment or soil, in which the organisms are known to do well. It might be the substrate where the organisms were collected, or in which they were cultured.

Statistical analyses. Comparisons of effect are made with the reference sediment/soil unless it is unsuitable, in which case the control sediment/soil is substituted. Analysis and interpretation follow standard methods described in other sections, and are exemplified in the document for polychaete worms (EC, 2001a). Single-concentration tests are limited to hypothesis testing (Section 7). If the tests involve several dilutions of test material, or spiked sediments/soils, analyses can produce point estimates, either ICp (Section 6) or EC50 (Section 4).

2.8 Reference Toxicants and Warning Charts

Key Guidance

- *Periodic tests with a standard ("reference") toxicant are intended to assess changes in sensitivity of organisms and precision within the laboratory.*
- *The laboratory's historic results are plotted on a warning chart. The new value is compared with the mean of previous results, and with the warning limits of ± 2 standard deviations. All calculations are based on logarithmic concentrations; failing to do that is a common mistake in Canadian laboratories.*

Tests with a *reference toxicant* are quite different in purpose and characteristics from the controls described in Section 2.7. They use a standard toxic chemical at known concentrations, to measure the relative effects on test organisms and are normally repeated over the months the laboratory is active. The intended purposes are (a) to detect any change in sensitivity of organisms over time, and (b) to

assess any fluctuation in the measuring technique of the laboratory. Those two causes of variation cannot be distinguished, especially since it is not customary to carry out the tests in replicate. The test is completely distinct from any toxicity test with samples, although it is often done at the same time. Environment Canada test methods for aquatic and terrestrial organisms require that reference toxicants be used periodically.

The use of a reference toxicant can also be called the use of a *positive control*. The common reference toxicants used are phenol, sodium chloride, or one of the metals (EC, 1999a). The results must be shown in a warning chart, to judge whether variation of results at the laboratory is satisfactory. The warning chart might be similar to that in Figure 2, although the points and labelling might be done by hand.

Figure 2 shows results for part of a series of reference toxicant tests with rainbow trout. The dates and timing have been changed from data obtained at a Canadian laboratory. The figure indicates three tests in each quarter of the year, i.e., monthly. The mean of the logarithmic EC50s is -0.027356. The antilog of that is the geometric mean of the EC50s, 0.94 mg/L, and that is shown by a line drawn across the chart. *Warning limits* were calculated as plus-and-minus two standard deviations, and plotted on the chart as horizontal lines, serving as visual indicators of any trend towards divergent results. The standard deviation was calculated as 0.15288 from the data of Figure 2, so two SD would be 0.30576. Adding and subtracting two SD from the mean yields logarithms of -0.33312 and 0.27840 as the warning limits, giving antilogs of 0.46 and 1.9 mg/L (plotted in Figure 2).

When a laboratory obtains a new EC50 for the reference toxicant, it is plotted on the control chart as in Figure 2. If it is within the warning limits, the result would be considered satisfactory. If the new EC50 is outside the warning limits, the laboratory should start an investigation of the causes for the variation. The new EC50 would then be incorporated into the "historic data" for all previous reference toxicant tests at the laboratory, and a fresh set of calculations made for the geometric mean and the warning limits. Those limits would apply to the next test of a reference toxicant. It is not difficult to set up a spreadsheet to carry out such sequential calculations in a proper fashion, and provide a plot similar to Figure 2.

All calculations for the warning chart must be done on a logarithmic basis of concentration, and the vertical axis of Figure 2 has a corresponding scale. The reasons for adhering to logarithms are given in Section 2.3 and the manner of calculating the mean and standard deviation is shown in Appendix F. If the arithmetic values of the EC50s had been used, with arithmetic calculations throughout, rather different warning limits would have been obtained. The average would have been 0.99 mg/L, somewhat higher than the proper value of 0.94 mg/L. The upper warning limit would have been 1.6 instead of 1.9 mg/L, and the lower limit 0.39 instead of 0.46 mg/L. The arithmetic range between limits would have been smaller, at 1.2 mg/L instead of 1.4 mg/L (Appendix F).

In Canadian laboratories, one of the most common failures in procedure is resistance by investigators to making logarithmic calculations. Indeed, many of the earlier methods documents published by Environment Canada treat this matter as a recommendation and not a requirement.

Some computer packages offer an option to calculate and plot a "control chart". Investigators should check that the program makes its calculations using the logarithms of concentrations. For example, CETIS (2001) will plot a chart with the mean plus-and-minus two standard deviations, but they are incorrectly calculated as arithmetic values of concentration. The same error is present in TOXCALC (1994).

2.8.1 Reasonable Variation

The magnitude (span) of the warning limits is obviously important. Narrow limits signify great precision in historical results from a laboratory. Accordingly, if that laboratory obtained an EC50 outside the limits, it would not necessarily indicate a serious problem in procedures at the laboratory, or a serious change in sensitivity of organisms. In fact, approximately 5% of the EC50s would be expected to fall outside ± 2 SD by chance alone. Conversely,

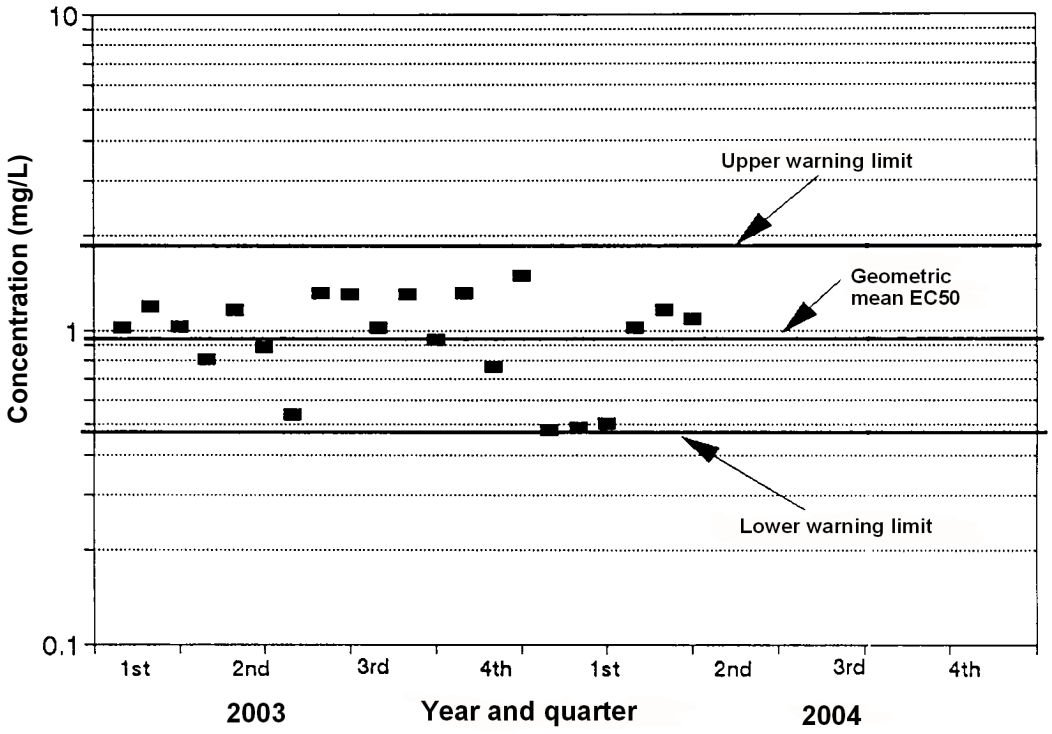


Figure 2 **A Warning chart for tests with a reference toxicant.** This shows real data from a Canadian laboratory, for aquatic tests with a reference toxicant. There is a fairly regular pattern of EC50s within the warning limits (some of them just within). The overall variation is slightly greater than might be considered desirable. The vertical axis and all calculations are based on logarithmic values for the EC50s.

a laboratory might have erratic historic results leading to wide warning limits; subsequent EC50s might lie within the warning limits but still indicate undesirable variation.

Accordingly, a second way of assessing variation can be considered, which might be called a "reasonable" degree of variation. This topic is separate from the warning limits described previously. Environment Canada has not set any firm definition of a reasonable degree of variation that might be expected in a set of repeated tests. However, it was suggested that a coefficient of variation (CV) no greater than 30%, and preferably ≤20%, might be reasonable for tests with reference toxicants (EC, 1990d) [13]. The same guideline for

[13] The coefficient of variation (CV) equals the standard deviation divided by the mean, customarily expressed as a percentage. That calculation of the CV is valid for arithmetic data. Accordingly, a known CV can be used to calculate the SD; simply multiply the mean by the CV (as a decimal proportion). This way of calculating CV is not valid for a set of logarithms. For lognormal data, the formula is CV = square root $[10^{(SD \times SD)} - 1]$. The SD in the superscript is the one calculated from the logarithmic data. The SD multiplied by itself is the variance, which could be used instead, as the superscript in the formula.

variation with reference toxicants has been suggested in the method for testing sediment with polychaete worms (EC, 2001a).

The guideline mentioned by Environment Canada was apparently based on the arithmetic means and standard deviations for sets of EC50s, and accordingly will have some degree of bias. However, an equivalent guideline can be calculated on a logarithmic basis of concentrations, and that has been done for this document (see derivation in Appendix F). The guideline or rule of thumb for reasonable variation would be this: **the value of the standard deviation calculated from the logarithmic data should not exceed 0.132, and preferably should not be higher than 0.0338**. These same logarithms can be applied to evaluate the SD of *any set of endpoints*. They correspond approximately to the previously mentioned arithmetic CVs of 30% and 20%, but avoid the possible distortion.

The guideline herein requires the mean and standard deviation of the set of endpoints to be calculated using logarithms of concentration. The calculated standard deviation (a logarithm) is compared with the value 0.132, and if the calculated value is equal or smaller, the variation in the set of endpoints is considered reasonable.

The suggested guideline for reasonable variation could be compared to the example data used in Figure 2. Those data have a calculated logarithmic SD of 0.15288..., so the observed variation in endpoints somewhat exceeds the "reasonable" guideline.

Some appreciation of this "reasonable" variability can be gained by creating a hypothetical scenario to compare with the situation shown in Figure 2. If the hypothetical set of EC50s had the same mean, but happened to have the "reasonable" SD of 0.132, the warning limits would be 0.51 and 1.7 mg/L (see Appendix F) [14]. Those warning limits for the "reasonable" hypothetical data would be somewhat narrower than the ones for the data shown in Figure 2.

If a set of hypothetical data had even less variation, with an SD equal to the "preferable" value of 0.0338, and the same mean as shown in Figure 2, the limits would be very narrow. Converted to arithmetic values for comparison with Figure 2, the limits would be 0.80 and 1.1 mg/L, which seems somewhat optimistic for the variation among repeated toxicity tests.

There would not be any constant relationship between this suggestion for assessing reasonable variation, and Environment Canada's long-standing warning limits of ± 2 SD for reference toxicants. The rule of thumb (or guideline) for reasonable variation would stay constant, but the warning limits would vary with the set of data.

2.9 Transformation of Effect Data

Key Guidance

- *For quantal data, a common and standard transformation uses probit or logit of effect to estimate the EC50.*
- *In estimating sublethal endpoints by regression, there is an assumption of normal distribution of residuals, and transformation can help achieve that. It can also simplify the relationship for use of regression. A major disadvantage is that transformation requires individualized weighting to compensate for altering the variances of the groups of observations; this requires statistical skill or advice.*
- *In hypothesis testing, if the data for effects fail to meet requirements for normality and homogeneity of variance, transformation might remedy that and allow analysis by standard parametric methods. This is recommended as the first option if hypothesis testing is desired and the data do not meet requirements.*

[14] Investigators who are rusty in doing calculations with logarithms might check their procedures by consulting the entry in the glossary, Section 2.3.5, Appendix D, or going through the arithmetic in Appendix F.

- *Logarithms and square roots are commonly used transformations. Arcsine square root is for quantal data; it and the reciprocal are not often useful.*

Transformation of results might assist either of the two broad approaches for analyzing data from environmental toxicity tests -- regression techniques and hypothesis testing. Both approaches have certain requirements for normal distribution of the effects data. If normality does not prevail, one option is to transform the data to make them satisfy the requirement. For regression, there can also be the additional pressure to transform in order to obtain a straight-line relationship for simpler analysis.

2.9.1 Use in Regression

For tests of acute lethality or other **quantal effects**, it is customary to transform the percent effects to probits or logits. Those transformations are suitable and advantageous for statistical analysis. The probits or logits generally eliminate the s-shape in a set of data (Appendix H), allowing a straight-line model and fewer parameters to estimate. These benefits are described in the following text under "**Advantage** ..." The conventional use of probits or other transformations for quantal data are further discussed in Section 4. (For concentration, logarithms are retained for analysis [15]).

Analyses of data from **sublethal tests** are moving towards more advanced techniques, notably nonlinear regression (Section 6.5). The construction of confidence intervals for the parameters of nonlinear regression models usually assumes that the residuals are normally distributed. Transformation would, again, be one possible approach for achieving this requirement.

Advantage of transforming: simplicity. One important principle of regression techniques for point-estimates is to keep the model simple, if that can be reasonably done. Transformation of the data can simplify the relationship and allow the use of a simple model. Although models can be created to fit a complicated relationship, the resulting equation will have many terms, and consequent loss of degrees of freedom, weak predictive power, and possibly widening confidence limits for the predicted endpoint (Andersen *et al.*, 1998). Statisticians stress this feature of modelling, e.g., "Thus simplicity, represented by parsimony of parameters, is ... a desirable feature of any model ... Not only does a parsimonious model enable the research worker or data analyst to think about his data, but one that is substantially correct gives better predictions than one that includes unnecessary extra parameters" (McCullagh and Nelder, 1989). It follows that transformations of the data could be beneficial, in allowing a simpler model.

The equation that fits exponential growth is a simple example of transformation (see Section 6.5.3).

$$Y = \alpha\beta^X \quad \Rightarrow$$

$$\log Y = \log \alpha + X \log \beta \qquad [\textit{Equation 1}]$$

Using logarithms, this equation can be transformed from a multiplicative relationship (upper line, Equation 1) to a linear one (lower line, Equation 1), allowing a relatively simple regression.

A common transformation to make proportional data fit a normal distribution with equal variances is taking the arcsine of effects (see Glossary and Section 2.9.3). These kinds of transformation seem to be an easy way of making the analysis simpler and reducing the "obstacle of calculating confidence intervals around nonlinear regression estimates ..." (Nyholm *et al.,* 1992).

Disadvantages for regression. The apparent advantage of transformations tends to be outweighed by some major complications. Although transformations might be intended to simplify the estimation of parameters, they can distort a real (mechanistic) relationship. For example, enzymatic reactions are mechanistically described by the nonlinear Michaelis-Menten equation. True threshold-concentration effects could also be distorted by inappropriate transformations.

---

[15] Use of the logarithmic values of concentration (and/or time) in the analysis simply retains the original units of the exposure, for basic scientific reasons (Section 2.3). Transforming log concentration to arithmetic values for a set of calculations, aside from being incorrect, would likely also introduce skew into the relationship and require a more complex model.

Sometimes, transformation can lead to highly biased estimates of endpoints, occasionally described as "fatal". The problems are discussed from a toxicological viewpoint by Christensen and Nyholm (1984) and Nyholm *et al.* (1992). They point out that transformation requires proper weighting to compensate for altering, to different degrees, the variances of the data-points. The weighting is specific to the data obtained, so there is no "cook-book" statistical package that can be applied. Weighting factors must be inversely proportional to the variance of the data as calculated for the original measurements (observations) at any given value of the independent variable X (or usually, logX). Even then, compensation by weighting might not be precise enough for irregular data, or for observations near the extremes of the dose-effect distribution. The weighting would also have to take into account whether the original variance was in absolute units or was proportional to the magnitude of the measured variable.

Such individualized statistical tailoring of data-sets is well beyond the limits that can be defined in EC methods documents for routine tests. Investigators should be aware of the pitfalls that can be encountered by transforming results to obtain a linear regression. Further, it is recommended that if investigators think that transformation might be advantageous, they should seek the advice of a statistician familiar with toxicity testing. It is possible that suitable statistical approaches are already available (see Section 6.5.8), or that statistical packages could become available in the future.

2.9.2 Use for Hypothesis Testing

The most familiar methods of hypothesis testing assume that results have a normal distribution. That assumption prevails for t-tests, ANOVA, and multi-comparison tests. Accordingly, data must be tested for normality before proceeding to the analysis (Section 7.3).

If a set of data does not show normal distribution, the investigator has three main options:

- use a more sophisticated parametric method that is appropriate for the data,
- transform the data to achieve normal distribution, or
- use a non-parametric method that has no assumption about distribution.

The most desirable choice is the first one, but this choice is seldom made because the methods are not known to most investigators, who are not statisticians. Historically, standard procedures came to be based on other approaches, because the more sophisticated parametric methods involved arduous calculations, but that is not an impediment since the advent of computers. This document does not give guidance on these more advanced methods, but provides some introduction to them in Sections 6.5.2 and Section 6.5.11 on GLIMs. It is to be hoped that future interchange with statisticians will make such improved methods available and workable for environmental toxicology.

The most favoured approach has been the second one listed, transformations to achieve normal distribution. This historical use of transformations has had the objective of producing data suitable for the statistical methods of earlier decades. It allows the use of known standard methods of analysis, with relatively simple procedures and readily available statistical tables.

The third listed approach, of non-parametric methods of analysis, has also become a standard modern approach, due in part to development and programming of standard methods for hypothesis testing in the USA. Non-parametric methods have customarily been used when parametric analysis is not valid. In many cases they are less powerful than corresponding parametric tests at distinguishing effects. Like parametric tests, nonparametric ones also have assumptions about the data such as independence of observations and homogeneity of variance, but they are generally more robust with respect to departures from those assumptions.

Advantages and disadvantages. If it is desired to analyze results by hypothesis testing, transformation can take measurements that digress from normality or homogeneity of variance, and change them into variables that conform to requirements for analysis by the familiar parametric tests. Another use is to

transform some sets of quantal data, to make them more suitable for hypothesis testing (Section 2.9.3).

Accordingly, suitable transformation is recommended if necessary, as the preferred option for data that do not satisfy requirements for normality and homogeneity of variance. Consultation with a statistician is advisable. The most serious problem is that transformation can be expected to change the inter-relationships within the data. The cautions of Section 2.9.1 must be considered.

If a satisfactory transformation is not found, the next option would be analysis using non-parametric tests.

2.9.3 Specific Transformations

The most frequently used transformations of measurements are the logarithm and the square root. Both can be effective if variance increases with the magnitude of the mean. **Logarithms** are useful if the effect tends to increase exponentially with increasing concentration and if variance is proportional to the square of the mean result for the treatment. This might happen with population growth, or weight, and transformation might make the variance independent of the mean. The preferred form, especially if some of the values are small or zero, is log (X + 1).

The **square root transformation** can also help to stabilize the variance. It is also applicable when the data arise as a series of counts (Poisson distribution), and the group variances are proportional to the means. Again, the preferred form includes a constant rather than a simple transformation, commonly the square root of (X + 0.5), where X is an individual measurement (Zar, 1999). Possibly superior to that is the slightly more complex transformation of the square root of X plus the square root of (X + 1).

The **reciprocal** transformation is seldom useful for quantitative data. The **arcsine square root** transformation is not recommended and is not relevant for quantitative data, because it is intended for binomial observations, such as percentages or proportions (Zar, 1999). Sometimes, however, an investigator might wish to analyze quantal data by hypothesis testing, and arcsine transformation could be useful and is suitable. Section 7.2.6 discusses the topic, and the application of arcsine is discussed in the Glossary.

Single-concentration Tests

Key Guidance

- *Single-concentration toxicity tests are typically used in surveys to assess contaminated sediments and soils, or monitor effluents. Their results can be used to judge compliance with regulations, using a fixed regulatory pass/fail criterion, without statistical analysis.*
- *Testing for statistically significant effect in single-concentration quantal tests (e.g., mortality) depends on the type of investigative program and its design. For a sample from one location without replication, testing could be done by a comparison with the control using Fisher's exact test, or "Finney's tables". For a single location with "field" replicates, (e.g., a survey of contaminated sediment or soil) results could be tested with Fisher's exact test.*
- *For a survey of several locations, with no replication and quantal effects, results would not be statistically testable. With field replicates, results could be assessed by logistic regression carried out by, or under supervision of, a statistician; ANOVA might sometimes be feasible.*
- *Quantitative single-concentration tests (e.g., effects of exposure to contaminated sediment on the weight attained by organisms) have different statistical methods. For sampling at one location with field replicates, results could be compared to the control with a t-test. Without replicates, results could not be tested statistically.*
- *For quantitative results at several locations, there are a number of approaches. Without replication, no statistical analysis is suggested. With field replication, ANOVA would be a first step if results were suitable. If the null hypothesis of no difference was rejected, analysis would proceed to one of several multiple-comparison tests. For ordered data (gradient expected), Williams' test would compare each location with the control. For unordered data, Dunnett's test would compare with the control, and the Dunn-Sidak test would be a second choice. For pairwise comparison (each location with each other location) Fisher's LSD is recommended, with Tukey's test as an alternative.*
- *For replicate field samples, and quantitative data requiring non-parametric analysis, in most cases it is recommended that the null hypothesis be tested before proceeding to a multiple-comparison test. If the data are ordered, locations should be compared with the control by using Shirley's test. Pairwise comparison of ordered data would start with the Jonckheere-Terpstra test, and proceed to the Hayter-Stone test if the null hypothesis was rejected. For non-ordered data, comparison with the control would start with the Fligner-Wolfe test or if it was not available, with the Kruskal-Wallis test. If the null hypothesis was rejected, the Nemenyi-Damico-Wolfe test would be applied, with Wilcoxon Rank Sum test as second choice, and Steel's Many-One test as an alternative. For pairwise comparison, the Kruskal-Wallis test of hypothesis would come first. The recommended multiple-comparison test is the Critchlow-Fligner-Steel-Dwass test, with the alternatives of Steel's Pairwise test, or repeated use of Kruskal-Wallis.*

Single-concentration tests are often used in environmental programs to monitor discharges for regulatory compliance, and to explore potentially polluted areas of sediments, soils, or surface waters. The tests, although not powerful, are efficient, meaningful ways to fulfil the exploratory purposes.

A program to monitor compliance of liquid discharges might simply use full-strength effluent, without any accompanying dilutions. Effects of the effluent would be compared with those in the most appropriate control that could be selected. Initial tests of potentially contaminated soils or sediments

are usually done by using an undiluted sample. Effects are normally compared with reference and control soils/sediments (see glossary).[16]

No statistical comparison to the control is necessary when a single-concentration test, such as measurement of acute lethality, is used under the regulations for Canadian metal-mining or pulp and paper mills. The test material would either pass or fail if mortality exceeded the allowable limit.

For some other single-concentration tests, standard statistical procedures are available for analyzing the results, and hypothesis testing is often used, with desirable replication of samples in the field. In analysis, the distinction between quantal and quantitative results must be maintained.

The common designs and choices of analytical methods are indicated in Figure 3, and discussed in the following sections. There can be many variations in single-concentration tests, to meet the needs of particular situations. For special designs which are not covered here, investigators should rely on the specific instructions of the method, consultation with a statistician, and any general principles outlined here.

3.1 Quantal Effects

Mortality is the most common endpoint in single-concentration tests, and the resulting data are quantal. A test might assess mortality of amphipods or midge larvae exposed to full-strength sediment, or mortality of rainbow trout in full-strength effluent. The upper left part of Figure 3 indicates the choices of statistical tests.

3.1.1 One Sample without Replication

Testing one sample without replication is common practice for a waste discharge being monitored periodically. By testing one unreplicated sample and a control, the number of dead can be compared using one of the tests outlined in the next paragraph. The comparison should be done as a one-tailed test of significance, because normally, the investigator would be concerned only with greater mortality in the test concentration than in the control or reference [17]. Because the tests are based on limited data, they can only be expected to detect relatively large effects.

The two suggested methods follow. Appendix G provides examples and references, although these tests of proportions are covered in standard statistical textbooks. In comparisons which use these methods, the null hypothesis is that the test concentration does not show an effect that is "worse" than the performance observed in the control, i.e., a one-tailed test mentioned previously. The procedures work whether or not the control shows reduced performance (e.g., some mortality).

- ***Fisher's exact test*** is recommended as a first choice, because it is indeed, an *exact* test. It is carried out with minimal calculations, in step-by-step selections and manipulations of the data laid out in a simple two-by-two table. A calculated value is compared with a critical value for Fisher's exact test, as provided in general statistical texts.

- **"Finney's tables"**. The procedure is a simple matching of the data with some diagrams, which immediately show whether the experimental effect is higher than the control performance. The diagrams are shown in Appendix G, but they are only for equal numbers in the test and control chambers, up to ten individuals. For larger or unequal situations, one might consult the source of these diagrams, the published tables of Finney *et al.* (1963). The book of tables is in some university libraries, but might be difficult to

[16] Subsequent tests with soils might establish a series of concentrations by diluting with clean soil, in which case ICps or EC50s might be estimated.

[17] A two-tailed test such as chi-square is used if the direction of the difference is not important, or cannot be assumed before starting the test. That would seldom be the case in lethality tests, which deal with greater mortality in the test sample than in the control.

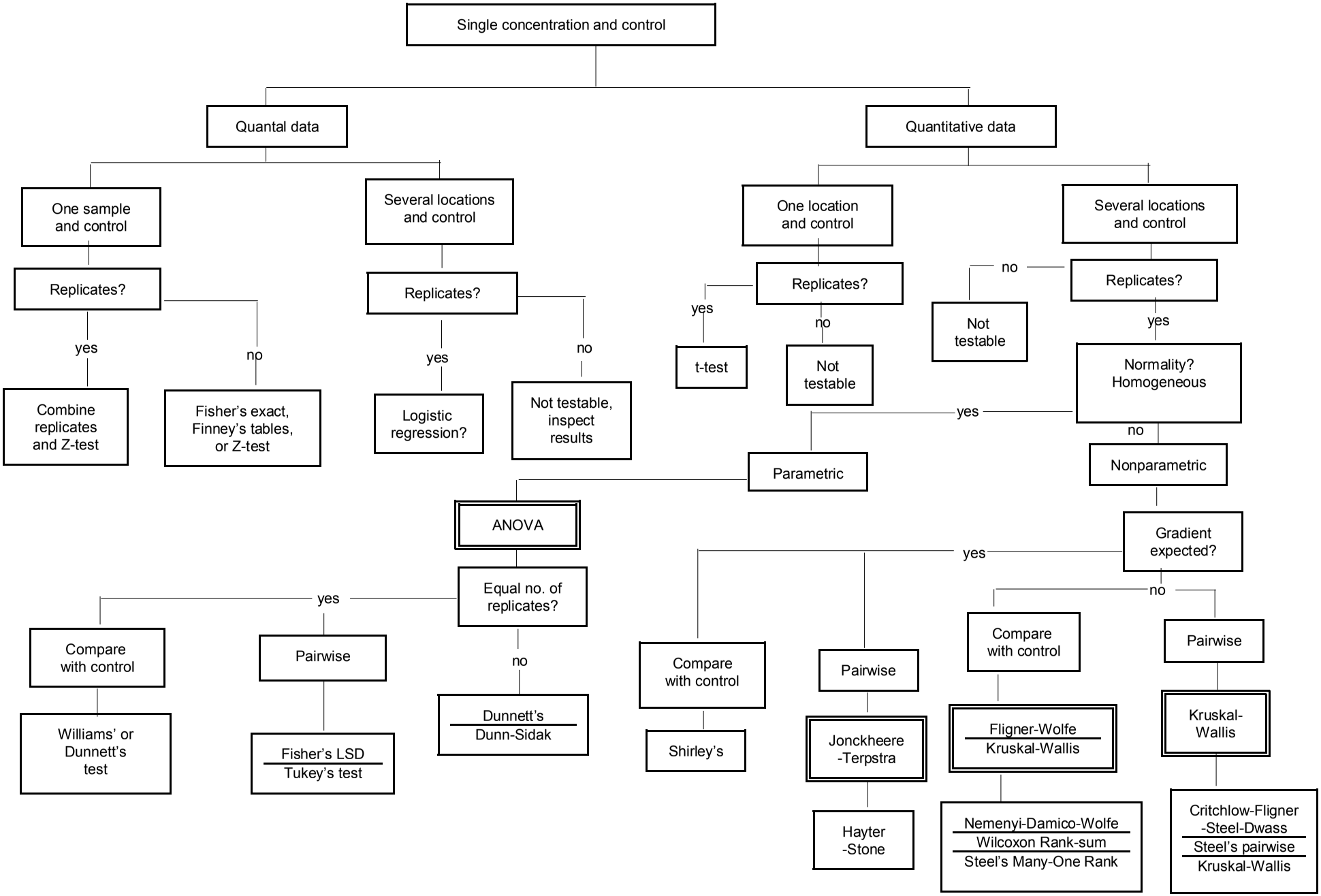


Figure 3 **Sequence of potential statistical procedures for various categories of single-concentration tests.** Some of these options might seldom be used. A box with double outline indicates a test of a null hypothesis; only if that is rejected, does multiple comparison proceed.

find. The tables represent a tabulation of p-values for comparing two proportions, conceptually similar to the t-test [18].

The ***Z-test*** is another way of comparing two proportions. It is not recommended here because the two tests listed previously are available. The Z-test is listed in most North American statistics texts[19] (e.g., Zar, 1999, p. 557), and an example is worked in Appendix G. A calculated value is compared with a critical value of "Z", which in fact, is found in tables for values of "t". The test relies on the normal approximation to the binomial distribution, which is poor for the small sample sizes that would be available in the comparisons considered here. The approximation is especially poor when observed proportions fall outside a range of about 0.4 to 0.6.

Neither Fisher's or Finney's method should be interpreted too closely. The exact probability value will usually be returned by the computer program for Fisher's exact test. Even if it were necessary to look up critical values of Z in tables, the investigator would be able to judge the approximate p-value. As a rule of thumb, the significance of p-values in the general range from 0.025 to 0.075 might be considered inconclusive. For important studies, further testing might be done, or a statistician should be consulted about other options. If alternative statistical tests were available, they would have to be selected to match the characteristics of the particular toxicity test.

[18] Two proportions can be compared, using the same ideas that underlie the comparison of two means. When comparing two means a *t-distribution* or a *normal distribution* is used to determine which difference between two means is statistically significant. Ever-smaller differences in means can be detected, as the sample size becomes larger and the variability associated with the means becomes smaller. The same approach can be used to compare two proportions, but the *binomial distribution* is used. The calculations are described in Zar (1999) and are somewhat tedious. It appears that Finney *et al.* (1963) made such direct comparisons of two proportions to construct their tables and the diagram in Appendix G.

[19] In European textbooks, the symbol Z signifies the standard normal variable, and its values are found in tables of normal distribution.

3.1.2 Replication at One Location

A program of testing at one concentration might sometimes use field replicates from a single location, i.e., several samples collected at the same time and place. This would be more likely in soil or sediment programs than in monitoring liquid effluents. There are no customary statistical procedures established, for using the full spectrum of data on quantal effects, but some options remain. In this situation, Fisher's exact test is still appropriate; however, the equality of the replicates should be tested (with Fisher's exact test) before pooling the data. If the test shows that the data cannot be pooled, the investigator is left with a serious question of why effects are significantly different at one location. Another possible analysis would combine the data from replicates and test the proportions with a Z-test, as listed in Section 3.1.1.

3.1.3 Multiple Sampling Sites

If single samples (e.g., sediments) from a number of locations were tested at one concentration with a control, the opportunities for statistical testing of the entire set of data are virtually non-existent. Usually, such a study would be exploratory. The results could be inspected for indication of strong effect, and further sampling and testing with replication could be done (see following text).

Some data-sets might lend themselves to special analyses, in consultation with a statistician. It is possible that outlier analysis might be applied to identify any effects that were more severe than in the control and low-toxicity samples (Section 10.2). If the locations constituted a gradient (e.g., upstream-downstream), a regression might be conducted to test for the gradient effect.

Subsamples of Each Sample. Single samples of sediment, soil, or liquid, from each of several locations, plus a control/reference, might be divided into subsamples and tested. That would represent "laboratory replication". For quantal effects, such data limit the options for statistical analysis (see following text). The laboratory replication gives an indication of the variation in the toxicity tests conducted in the laboratory and the homogeneity of the sample. If the subsampling variation were very low, the replication might assist in distinguishing among the field samples. For example, if

subsampling variance was close to zero, it would indicate good homogeneity of the samples and precise results of toxicity testing; differing toxicities of the field samples would be noticeable. However, variation in the field sampling at a given location would remain unknown, so the subsampling would not produce any power to judge differences among locations. For this reason, laboratory replicates are not particularly recommended unless it is specifically desired to assess variation within the laboratory. Generally, it would be more useful to focus the extra effort on field replicates (see Section 2.5.2). Conclusions from statistical analyses with laboratory replicates should be made cautiously and their meaning must be clearly stated. Otherwise, the statistical findings might be misinterpreted, with a mistaken inference that any detected differences resulted from the different field locations.

Field Replication. If field replicates were taken, i.e., several samples at each location, useful statistical analysis becomes feasible, even for quantal data at one concentration. A possible approach would be logistic regression (Section 6.5) carried out by a statistician or a toxicologist well versed in statistics. The regression would be "categorical", i.e., based on Control, Location 1, Location 2, etc., rather than the familiar regression on a continuous independent variable such as concentration. The approach of logistic regression might be particularly fruitful if a gradient of effect was expected (e.g., at successive locations "downstream" of a pollution source).

3.2 Quantitative Effects at One Location

An example of a single-concentration test for quantitative effects would be measuring the average weight of midge larvae after exposure to a sample of undiluted sediment, compared to weight of midges exposed to a reference or control sediment (EC, 1997a). Exploratory tests might conceivably run single test containers, although definitive tests would use field replicates. The extensive branching of choices is shown in the right and lower parts of Figure 3.

Without replication. If there was only one sample tested, and one control or reference material, without any replicates, results could be not be compared by any statistical test.

Replication and comparison by t-test. In a quantitative test with replication for the test material and for the control or reference material, a standard ***t-test*** would be suitable for statistical analysis. Here again, the investigator would be looking for smaller size in the test material, so the critical value for the t-test would be for a one-tailed test. The procedure for t-tests is commonly provided in statistics texts and in software programs such as TOXSTAT.

As previously discussed (Section 3.1.3), if the replicates were subsamples of a single sample ("laboratory replicates"), the conclusions from statistical testing would only reflect within-lab variation. No conclusions could be drawn about differences in the outside world, for example, whether the sampling location differed from the control location. If field replicates were used, however, the conclusions would apply to the real world at that time/place.

The ***t-test*** can be applied to most sets of data. It functions for unequal numbers of replicates in the test and control. Strictly speaking, the t-test assumes a t-distribution and equal variances in the two groups. If there was doubt about those assumptions, the t-distribution could be tested by a quantile-quantile plot, or if the sample size was greater than about 30, by a test for normal distribution. Homogeneity of variance could be tested by ***O'Brien's, Levene's***, or ***Bartlett's test***, or the ***F-test*** (Section 7.3.1) [20]. However, the t-test is fairly robust, especially if sample sizes are equal or nearly equal in test and control, and if the numbers are not too small. Various modifications are available, and CETIS offers the *paired-sample t-test*, the *equal-variance t-test*, and the *unequal-variance t-test*.

---

[20] The F-test is the last choice, but if used, the method is found in all statistical textbooks, which usually provide tables for critical values of F. If there are four replicates and each has an average weight of surviving organisms, the variance is calculated from the four means, giving a variance for the test material and another for the control. F is the ratio between the greater and lesser variances. The degrees of freedom are one less than the number of replicates in each case. If the t-test is invalid because the variances are not equal, a modified formula for the t-test would be used. Worked examples are shown in an appendix of USEPA (1995).

3.3 Multi-location Quantitative Tests

In another kind of single-concentration test, samples from several places are tested at the same time, using the same procedure, and the same control or reference material. This is commonly done with samples of soil from various places around a contaminated site, or sediment from several different places within a harbour, in order to delineate any zone of high contamination. A thorough guidance document on sampling and replication for sediments is available for advice and is recommended (EC, 1994). There is also specific advice for individual sediment-testing methods such as that for polychaete worms (EC, 2001a).

Field sampling of sediment at different locations is used as an example here. Comments are for tests with quantitative effects such as weight of organisms.

Productive statistical analysis of sediment samples from several locations, requires separate samples to be collected at each location (i.e., *field replicates*). The manner of replication is covered in Section 2.5. For hypothesis testing, an alternative that is of no use in distinguishing locations is one sample from each station, divided later into *subsamples* (so-called "laboratory replicates"). Testing would provide only limited information on whether a particular sample was different from another particular sample. It would *not allow testing a hypothesis of no differences among the locations (sampling stations)* (see Section 2.5).

Special case for gradients. If a gradient of decreasing effects is expected at a series of sampling locations which extended out from a source of pollution, regression can be used as a form of hypothesis testing. The null hypothesis is that *no gradient exists*. The alternative hypothesis is that a gradient of effects exists with increasing distance from the source. Selection and use of an appropriate regression technique requires guidance from a statistician. Replicates are not necessary for this analysis; however, field replicates allow lack of fit to be tested and also make the regression analysis more powerful in statistical terms. Subsamples ("laboratory replicates") could be used by a statistician to reduce the error variance, but sampling effort should be focused on field replicates.

3.3.1 Parametric Tests

If the sampling stations can be meaningfully ordered into a gradient, the immediately preceding comment on gradients applies, and further guidance is given in the following paragraph. Without expectation of a gradient, hypothesis testing could be done if field replicates were taken.

For hypothesis testing, the choice for statistical analysis is analysis of variance, if results met requirements for parametric analysis (Section 7.3). If each sampling station is compared with a reference or control material, the ANOVA would be followed by ***Dunnett's test*** (Section 7.5.1), and this sequence is recommended here. Some old software programs might require equal numbers of replicates for Dunnett's test, but recent programs escape this limitation (see Appendix P.4.2). ***Williams' test*** might be used instead of Dunnett's test, if it is clear that there is a gradient of effects such as at a series of locations successively downstream from a source of pollution, and if hypothesis testing is used. Williams' test would compare effects at each location with those at the control location, but would take the ordered nature of the locations into account, providing a more sensitive analysis (see Section 7.5.1).

Conceivably, the investigator might wish to know which sampling locations were different from which others. Such a situation might be several field samples from each of one or more locations upstream of an effluent discharge, and similarly from a number of locations downstream, all tested at full strength. The investigator might wish to make pairwise comparisons within a larger survey of locations, such as whether the downstream location showing the "best" recovery could be distinguished from the upstream station. To make such an evaluation, an ANOVA could be followed by ***Fisher's Least Significant Difference (LSD)*** or by ***Tukey's test***.

The LSD is useful for paired comparison within a larger set of data because it is relatively easy to carry out, and can be extensible to cases with unequal replication. The LSD is not commonly found in computer packages for toxicity, but advice on using it is given in Section 7.5.1. Other advice on parametric multi-comparison tests is given in Appendix P, Section P.4.

Unequal replicates. As mentioned previously, Dunnett's test handles unequal numbers of replicates, in the modern statistical software packages which are most likely to be found in laboratories. Old packages for toxicology might have only the version for equal numbers. If a program for unequal numbers is not available, there is a modification which could be applied. It is explained in Newman (1995) and examples are worked in USEPA (1995). The other options for unequal numbers of observations are the ***Dunn-Sidak test*** or the ***Bonferroni-adjusted t-test*** (see Appendix P, Section P.4).

3.3.2 Nonparametric Tests

If results from multi-sample toxicity tests did not meet the requirements of normality and homogeneity of variance, nonparametric tests would have to be used. The options are shown in the lower right portion of Figure 3. Relevant comments and more details are provided in all sub-sections of Section 7.5 as well as comments on the availability of tests.

One branch of testing would prevail if an order or gradient is expected in the results and if each location is going to be compared with the control. ***Shirley's test*** could be used to make those comparisons with the control (Shirley, 1977). If an order is expected, and a pairwise comparison is desired (each location with each other location), the ***Jonckheere-Terpstra test*** could be used to test the null hypothesis of no differences (Jonckheere, 1954). If the hypothesis is rejected, testing proceeds to the multiple-comparison test of ***Hayter and Stone*** (1991).

Another branch of testing would apply if no order of concentration or effect is expected in the set of multi-sample test results. In such a case, the effects could be compared with those of the control, by testing the null hypothesis of no effect, using the test of ***Fligner and Wolfe*** (1982). If that was not available, the test of ***Kruskal and Wallis*** (1952) would serve the same purpose. If the null hypothesis is rejected, testing would proceed to a multiple-comparison test. The first choice would be the ***Nemenyi-Damico-Wolfe test*** (Damico and Wolfe, 1987). Alternatively, the ***Wilcoxon Rank-sum test*** would be second choice, or ***Steel's Many-One Rank test*** could be used (Steel, 1959).

A pairwise comparison (each location with each other) might also be desired if there was no expected order of effects. First, the null hypothesis (no effect of location) would be tested using the Kruskall-Wallis test. If a difference was concluded, the ***Critchlow-Fligner-Steel-Dwass test*** (Critchlow and Fligner, 1991) would be used to identify the difference(s). Alternatively, ***Steel's Pairwise test*** (Steel, 1960) could be used for balanced data (equal number of replicates), or the Kruskall-Wallis test could be used again, this time as a multiple-comparison test for unbalanced data.

Quantal Tests to Estimate ECp

At the end of a quantal toxicity test, each organism either shows or does not show the defined effect. The effect is *binary*: an earthworm dies or lives, an ovum is fertilized or remains unfertilized, a fish shows an avoidance reaction or does not. Binary and quantal are synonymous. Thus, most quantal tests are based on the proportion of organisms that showed the effect, after exposure to a fixed concentration of test material and a defined period of time.

Quantal results follow a *binomial distribution*, which determines the choice of appropriate statistical tests. An investigator seeking further background in a statistics textbook should look for chapters or sections on binary data and binomial distribution. Collett (1991) describes the methods of analyzing binary data, and points out that the familiar techniques of analysis of variance and simple linear regression, in the forms used with continuous (quantitative) data, are not suitable for direct use with quantal data (see end of Section 4.3). There are well-established methods for fitting models to quantal data, but methods for checking the fit are less well established. (While assimilating the good statistical advice of Collett, readers should be alert for statements on toxicological matters which might appear to be misleading at first glance, as explained in Section 2.3.1.) Other relevant statistics texts are Finney (1971; 1978), and also Ashton (1972) who focuses on linear logistic modelling, particularly suitable for data from quantal toxicity tests. Hosmer and Lemeshow (2000) is a more recent text on logisitic regression. Fleiss (1981) covers some aspects such as contingency tables.

Most quantal tests in environmental toxicity are based on acute lethal effect. Although such tests do not estimate a "safe" concentration, they have a long history in the development of environmental toxicology, and a large base of results has been accumulated. Quantal tests continue to be commonly used, particularly in regulatory testing, perhaps because they often use well-known species such as rainbow trout. These tests have useful features such as rapidity, reasonable economy, a clear endpoint, and an effect that is obviously deleterious. The tests can compare relative toxicities of materials or sensitivities of species, and can provide initial explorations of toxicity or monitoring of changes in an effluent [21]. In conjunction, there are well-established methods for statistical analysis. Some sublethal tests are also quantal, and use the same analytical techniques.

The general pattern of analysis is reasonably straightforward (Figure 4). Logit or probit regression (frequently called "probit analysis") is recommended for routine use if the data are suitable, because a long history of use means that well-tested and convenient analytical programs are readily available. If the data do not meet the requirements for probit or logit regression, there are alternative methods, and although theoretically less desirable, they are capable of handling the data commonly encountered (Section 4.3).

[21] Lethal tests are not necessarily inferior to sublethal tests; sometimes they are precisely the tool needed for investigations. An example of using lethality to explore complex scientific topics, is the powerful development of *Quantitative Structure-Activity Relationships* or *QSARs*, i.e., relationships between chemical structure of substances, and their toxicity to aquatic organisms. Massive research programs have used lethal tests to define a series of complex QSARs, so that chemical configurations can be used for efficient predictions about hazardous new substances with similar chemical structure (Broderius, 1991; USEPA, 1994e).

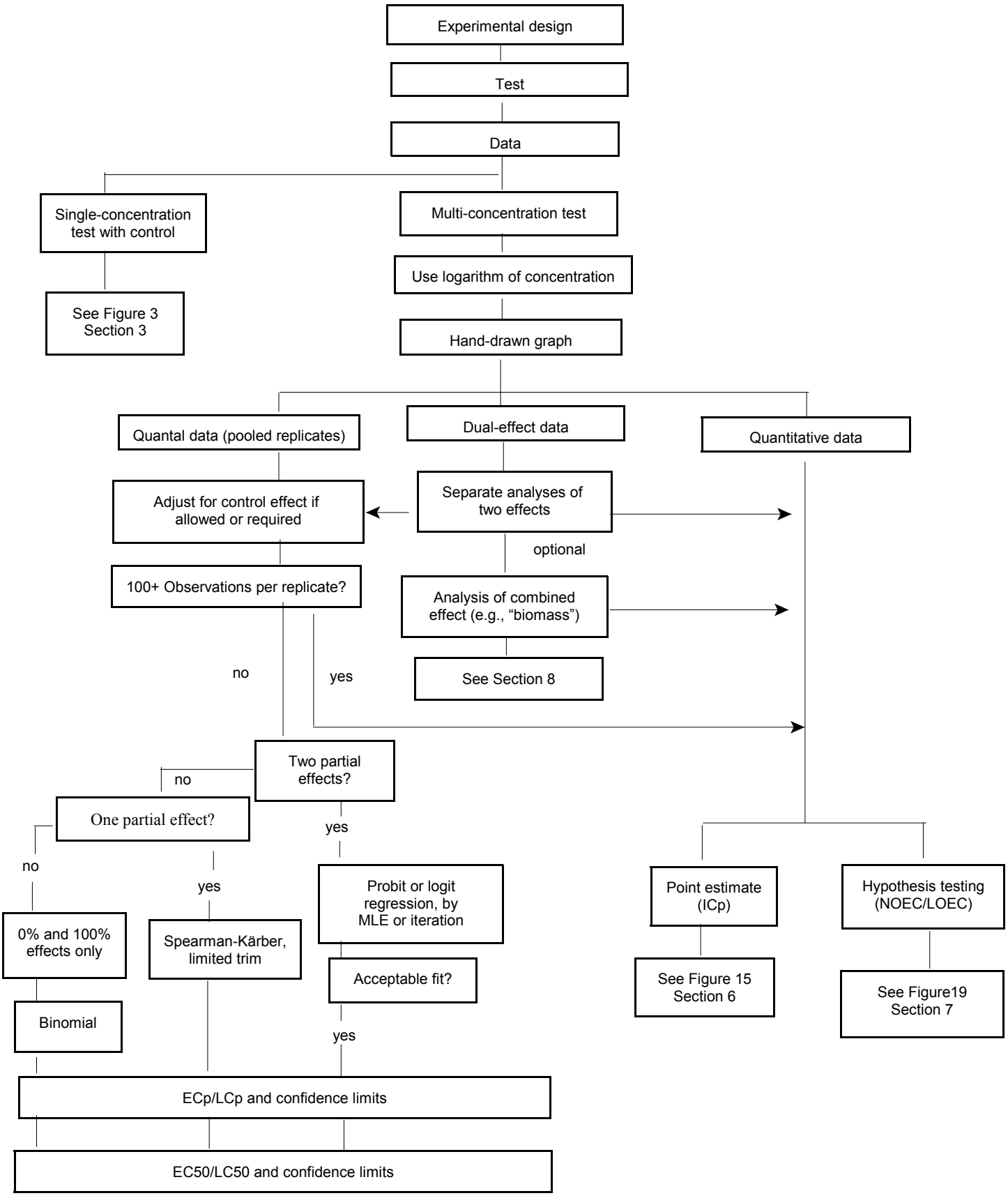


Figure 4 Sequence of analytical procedures for quantal tests. Quantal procedures extend downwards towards the lower left of the flow sheet.

Comments on the steps for statistical analysis of quantal data are given here in Section 4. Procedures are quite different for quantitative data, discussed in Sections 6 and 7. Certain tests with large numbers of quantal observations, can be analyzed by quantitative methods (Section 6.1.1).

4.1 The Endpoints of Quantal Tests

Key Guidance

- *In a quantal test, each organism either shows an effect or does not show it. The effect could be lethal or sublethal (e.g., immobilization).*
- *For a multi-concentration quantal test, the endpoint is the* Effective Concentration, *usually the* median Effective Concentration (EC50). *Lethal tests are a sub-category, and the usual endpoint is the* median Lethal Concentration (LC50). *The exposure time must be given, e.g., the 96-h EC50.*

In quantal tests, separate groups of organisms are exposed to one of a series of fixed concentrations, for a fixed time. It is desirable to have an equal number of organisms at each concentration, and the duration of exposure must be the same. The observations are the number of affected organisms at each concentration (e.g., number dead). The proportions affected allow suitable statistical analyses. (A background of practical analyses for proportions is given by Fleiss, 1981).

The endpoint of a quantal test is the *Effective Concentration for a toxic effect on a specified percent of test organisms*, the ECp. The chosen percentage *(p)* is usually 50%, i.e., the *median effective concentration*, expected to cause an effect in half of the organisms. In everyday terms, this is an estimate of the concentration that would just affect the "typical" or "average" organism, an endpoint of some validity. An additional reason for choosing 50% effect is that confidence limits are at their narrowest. They become wider as distance from the median increases, and so the limits would be very wide if extremely low or high values were chosen for percent effect (Section 4.2.4). There is currently some demand for estimates of EC25 or EC20, and those endpoints can also be estimated by some analysis programs (Section 4.2.5).

An exposure time must always be stated for an ECp, for example "the 96-h EC50". Quantal tests are commonly associated with acute exposures. The EC50 for viability of salmonid eggs after a 7-day exposure, for example, is an acute test because it represents a small fraction of the organism's life. Less commonly, a quantal test could be chronic, such as mortality among fish after months of exposure.

The term ECp applies to any quantal effect, lethal or sublethal. A commonly used sub-category is the *lethal concentration* (LCp, almost always LC50). In the following text, ECp or EC50 will be used as the more general terms that include LC50 [22, 23].

Confidence limits should be reported for each ECp (see Section 4.2.4).

4.2 General Procedures for All Methods of Estimating ECp

Key Guidance

- *The EC50 cannot be estimated by any method if there is not an effect ≥50%*

[22] Sometimes it can be difficult to determine whether an animal is dead, particularly for invertebrates. A suitable endpoint can be the EC50 for immobilization as in Environment Canada's test with daphnids (EC, 1990b). That endpoint is ecologically meaningful and should be accepted; it could well be used for other types of organisms.

[23] LC50, EC50, ICp, IC25, etc. have the grammatical status of nouns. There is no need to write "LC50 value" or "ICp estimate", in fact such expressions are redundancies. Each acronym is simply a short form of the full words, and the sentence structure should fit the full words. The operating word that comes from EC50 is "concentration" which is already a noun. One would not write "concentration value", and similarly it is incorrect to write "EC50 value". The most glaring example of this mistake, occasionally seen, is "LC50 concentration", signifying "median lethal concentration concentration".

in at least one concentration. The EC50 can be estimated if there is a zero effect at one concentration, but effects at all higher concentrations are ≥50%, if a logical linear effect is evident. An estimate of the EC50 is more reliable if partial effects bracket that endpoint. However, an EC50 can be interpolated from 0% and 100% effects at successive concentrations; it might be a precise estimate if the concentrations are close together.

- *In any estimate, information from concentrations causing no effect or complete effect should be used to help establish the position and slope of the dose-effect relationship, but only one zero effect and one 100% effect can be used.*
- *Concentrations must be plotted on a logarithmic scale to maintain the scientific assumption that was made in selecting concentrations. This scale usually eliminates skew, for easier visual comprehension of fit. Plotting the percentage effect on a logit or probit scale usually completes a transformation to a straight line rather than a sigmoid distribution.*
- *For estimation of endpoints by computer programs, there must be checks that observations have been correctly entered, and that the output of the program appears reasonable. One such check would be a hand-drawn graph of percent effect against log concentration, and examples are shown. The graph and its estimate of ECp should be compared to those produced by the computer.*
- *Endpoints such as EC50s are normally calculated as logarithms, then converted to arithmetic values of concentration for ease of comprehension. Before any averaging or other mathematical manipulation of EC50s is done, they must be converted back to logarithms. Time must also be handled in terms of a logarithmic scale.*

Certain general rules apply for all methods of estimating ECp, and they should not be circumvented. Computer programs do not necessarily guard against violations that could cause erroneous analyses.

- The data are combined for replicate containers at a given concentration [24].
- If an effect ≥50% is not achieved in at least one concentration, the EC50 cannot be estimated. (Of course, the EC50 can be said to be higher than the highest concentration tested.) Extrapolation must not be done from below 50% effect, to estimate a concentration that would cause 50% effect. It is possible that effects of 50% or higher might never occur at higher concentrations, e.g., a toxic chemical might reach its limit of solubility and fail to increase its toxicity further, or the remaining organisms might be *tolerant* of high concentrations.

An investigator must have secondary methods of analysis available, because many sets of results do not have the two partial effects required for logit or probit regression. At Environment Canada, laboratory staff have estimated that up to 90% of standard regulatory and monitoring tests "... result in either one or zero partial mortalities" (Doe, 1994) and therefore cannot be processed by probit or logit regression. Similarly, APHA *et al.* (1992) give an example that "... out of 60 acute [aquatic] toxicity tests performed, only four (7%) produced results that met the assumptions and data requirements of probit regression". It is often very important for monitoring purposes, to have estimates of EC50 and confidence limits that are reasonable, even if they are not perfect from a statistical point of view. The secondary methods will usually provide the reasonable estimate. Often, it is not feasible to repeat the test for a more precise or defendable result, because the sample of test material is either used up, or too old.

---

[24] If the results were hand-plotted to make a graphical estimate of EC50, the replicates could be kept separate to provide a visual impression of variation. The customary computerized methods of estimating ECp use pooled replicates. Possibly, future mathematical systems for analyses might be able to make use of data from separate replicates, but currently, few software packages are capable of correctly using this information.

4.2.1 Effects of Zero and One Hundred Percent

An estimate of ECp can usually be regarded as more reliable if the data show a partial effect below the ECp, and another partial effect above it. Nevertheless, an EC50 can be interpolated with no partial effects, if one concentration causes zero effect and the next highest concentration causes a complete (100%) effect (see Binomial method, Section 4.5.7). Indeed, such an all-or-none test might provide an excellent approximation of the EC50 if the concentrations were reasonably close together. The following guidelines are for use of zero and complete effects.

- It is permissible to estimate an ECp (e.g., EC25), from data which include a zero percent effect, but no partial effects at the chosen p% or lower. There must be a consistent pattern of effects above p%, compatible with a linear relationship, and the fitted line must describe a statistically significant proportion of the total variability [25]. Some authorities and computer programs might have stricter requirements for estimating an ECp [26]. The recommendation here, however, is that a test need not be disregarded for lack of partial effect below or at the chosen p%.

- If a certain concentration results in no effect, that information should be used in fitting the line. Similarly, an effect of 100% should be used. Those observations have low weight in fitting an effect-concentration line, but they help to establish the slope.

- If, however, successive concentrations yield a series of 0% effects or a series of 100% effects, only the "innermost" of the series should be used in estimating the EC50 (Ashton, 1972). In other words, the one that is used should be the highest of the successive concentrations that yielded 0%, or the lowest concentration that yielded 100%. In each case, the concentration (and effect) to be used is the one that is "closest to the middle" of the distribution of data. Use of only one 0% effect and/or one 100% is important for computerized analyses. If the investigator enters more than one successive 0% or 100% value, programs endeavour to use the additional value(s), change the slope and position of a fitted line, and so produce somewhat deviant estimates of EC50 and confidence limits. The solution to this problem is not to enter the "extra" values into the program. This is an important point and a common mistake. (These comments about successive zero percent effects do not, of course, apply to a control.)

4.2.2 Logarithmic-probability Transformation

In choosing the exposure concentrations for a test, an investigator is almost certain to follow the usual practice, by selecting them from a geometric/logarithmic series. That is a tacit admission that log concentration has been adopted as the appropriate dose metameter, and therefore is the appropriate base for subsequent statistical analysis, as explained in Section 2.3. Once calculated, the endpoint itself should be regarded as a logarithm. However, an endpoint such as EC50 is usually converted to an arithmetic value, to assist in our everyday comprehension of the numbers. A practical benefit of log concentration is that it usually eliminates *skew* in the plotted data (Figure H.1 in Appendix H).

Similarly, biological time is best treated as a logarithmic phenomenon (Section 2.3.6). Hence, logarithms of both time and concentration are used in toxicity curves (Section 5.2) and log time should be used in calculations, if median effective times are to be estimated (Section 5.1).

---

[25] The pattern should be demonstrated by plotting a graph. The chi-square should not exceed the critical value when a line is fitted by probit regression, a condition that applies to all tests (see Section 4.5.4).

[26] The computer program of Stephan *et al.* (1978) for probit regression (Section 4.5.3) requires two partial effects, as in all probit programs. In addition, the program requires either (a) one or more effects below 50% and one or more effects above 50%, or (b) an effect at 50% and at least one other, either below or above 50%. These are reasonable requirements, although slightly stricter than the present recommendations of Environment Canada.

Some computer programs might estimate an EC50 from inadequate data, but it should not be accepted unless the requirements of an Environment Canada test method are met. For example, the computer program of Hubert (1987) will provide estimates from two effects less than, or two effects greater than 50%. The former is not acceptable to Environment Canada, because there would be no data to prove that the effects would ever reach 50%. The latter (two effects >50%) would only be acceptable to Environment Canada if there were a zero effect at some lower concentration.

The use of probits to represent percent effect originated as a way to produce a straight-line relationship for the data. Empirically, the probits usually straighten a sigmoidal distribution of the effect-data, which was convenient in the pre-computer era because a straight line was easier for analysis. The practice has carried over into modern computer programs. Probits gradually "stretch" the vertical scale, for effects that are further away from the 50% level (as represented on a graph, see Figure 5 and explanation in Appendix H).

The combined log-probit plot produces a straight line, from what is really a cumulated log-normal curve (Appendix H; Buikema *et al.*, 1982; Chapter 1 of Rand and Petrocelli, 1985).

For hand-plotting, it is convenient to purchase "logarithmic-probability" graph paper and simply plot the arithmetic values. If log-probit paper is not available through a commercial supplier, the blank graph in Appendix I could be photocopied.

The descriptions and examples in the next section refer to probits for simplicity, but logits could be used, and the same general comments apply. The only exception would be that log-probit paper can be purchased, but not log-logit.

4.2.3 Estimate of EC50 by Hand-drawn Graph

Preparing a quick hand-drawn graph of results should be the initial step in gaining a general impression of the data and their resultant EC50. A working group of statisticians and toxicologists (OECD, 2004) agrees. They describe a "typical data analysis" and list the steps: "First, the data are plotted and visually inspected." Accordingly, graphical estimates are described first, and they conveniently illustrate some concepts and difficulties. To some extent, a graph explains what is being done by a program of computer analysis. Figure 5 shows examples using representative sets of data.

Most Environment Canada methods for quantal tests, and a guidance document of the USEPA (2000a), recommend an eye-fitted line to estimate an approximate EC50, to check the reasonableness of a computer estimate. A conscientious investigator should always make a hand-plotted graph, in order to apply the most useful assessment of validity -- common sense. Plotting might reveal an irregular pattern of effects which should not be forced into a standard mathematical analysis. A plot is especially needed if the Spearman-Kärber method of statistical analysis is being used (Sections 4.4 and 4.5.6). The hand-drawn graph can provide confirmation or warning, but it does *not* provide a definitive reportable endpoint.

Some investigators protest the need of constructing a hand-drawn graph in these days of splendid computer graphics, but the pencil and graph paper retain their importance. An error in entering the data on the computer would be reproduced in the computer's graph as well as in the mathematical estimate of EC50, and the agreement of the two would fail to detect the input error [27]. Investigators should, indeed, make use of the most modern and powerful computer programs available to them (such as maximum likelihood estimates). But from a practical point of view, a quick check by a hand plot might be the best way to remedy problems of incorrect reporting, which have been evident in previous EC testing programs. Some errors can be rather fundamental and simple. New workers might need time to develop skills in toxicity testing and statistical analysis. Data managers might enter test results without a good understanding of the analytical program, or whether its output was reasonable. Hand plots help to remedy those situations.

A computer-drawn graph should be compared with the existing hand-drawn graph. Alternatively, the results of statistical analysis might be plotted over the raw data or alongside the eye-fitted line, as a visual check. Any appreciable discrepancy should be investigated and resolved. Some examples of

---

[27] Most laboratories will have a quality-assurance program with an independent review of data that should detect any errors in data entry. If done rigorously, such a check can fulfil one function of a hand-plotted graph, but does not replace it. Computer programs can produce peculiar estimates from some sets of data, and we human operators tend to accept the output at face value. One EC laboratory reported a big discrepancy between the hand-drawn graph and the output of a newly purchased computer program. The program itself proved to be the problem, not the entry of data (reported by K.G. Doe, EC, Moncton).

graphs are given in this section, with advice on fitting lines by eye. With practice, those lines will produce estimates of the EC50 that are within a few percent of the computer estimates, thus serving as the reality check that is desired.

As mentioned, concentrations are plotted on a logarithmic scale, with percentage effect on a probit scale (Figure 5). Because the probit scale never reaches 0% or 100%, extreme values are plotted with an arrow as in Figures 5A and 5B. The arrow indicates that the true values lie somewhere beyond the 2% and 98% values which were the arbitrary limits on this graph paper. Despite their low weight, one 0% and one 100% effect should be plotted if available, because they sometimes help to "anchor" a line which has few data. Again, for a series of successive zero or complete effects, only the one closest to the centre of the distribution should be plotted.

In fitting a probit line by eye, a transparent ruler should be moved and rotated, in an effort to minimize the *vertical* distances between the observed points and the fitted line. At the same time, the points should be mentally weighted. Points closest to 50% effect should be given the most weight, and those at or near 0% and 100% should be given the least. As a rule of thumb, most weight should be assigned to those points between 16% and 84% effect, which are within one probit of the median. A value of 10% or 90% has about half the weight of a point in the 40–60% range. At 3% or 97% effect, the weight of a point becomes only about one-quarter of a value near the centre.

If in doubt as to where to place the line, a conservative approach is to decrease its slope thus implying more variation. As the slope of the line becomes lower, the confidence limits of the EC50 become wider.

Once the line is fitted, it is a simple matter to note the intercept with 50% effect, and follow down to the EC50 on the concentration axis.

Lines which might be eye-fitted to the examples in Figure 5 are discussed in the following examples. It will be convenient to make some comparisons with lines calculated by the formal statistical methods of probit regression, even though the mathematical methods are covered in Section 4.5.

Example A (Figure 5A). There is not much doubt where the fitted line should go. The observed data-points line up well, and most people would select a line very close to the one shown. That line is essentially the one calculated by probit regression, and its fit is good since the chi-square value is relatively low (chi-square = 1.11, Table 2). The calculated 95% confidence limits were taken from Table 2. The limits are narrow, as would be expected with a consistent set of data and a probit line with a high slope.

Example B (Figure 5B). An investigator might well fit the line shown, and it is essentially the same as the one estimated by computerized probit regression. There might be temptation to use a higher slope, to pass closer to the extreme values at 0% and 100%. However, this is a good example to show the lower weight that is assigned to extreme values; the line is strongly influenced by the three central points. The two extreme values do have a small effect, however, otherwise the calculated line would have had a lower slope to pass closely alongside all three central points.

Example C (Figure 5C). Most people would probably consider the dashed line to be a reasonable fit. It comes close to bisecting the upper and lower groups of points, and comes close to minimizing the vertical distances between the line and the points. The eye-fitted dashed line would estimate approximately the same EC50 as that calculated by probit regression on computer (solid line). It might seem puzzling that the calculated probit line goes to the right side of both upper points. Apparently the calculations produced a lower slope to fit the overall trend of all points, acknowledging the appreciable variation in this set of data (a relatively high chi-square of 3.5, Table 2).

Example D (Figure 5D). Such variable data as these might well be encountered in testing. The dashed line could be a reasonable choice for an eye-fitted line. It is less than perfect in paying too much attention to the 10% value which carries a low weight, and leaving a large vertical distance above the 50% value on the right side, a value that

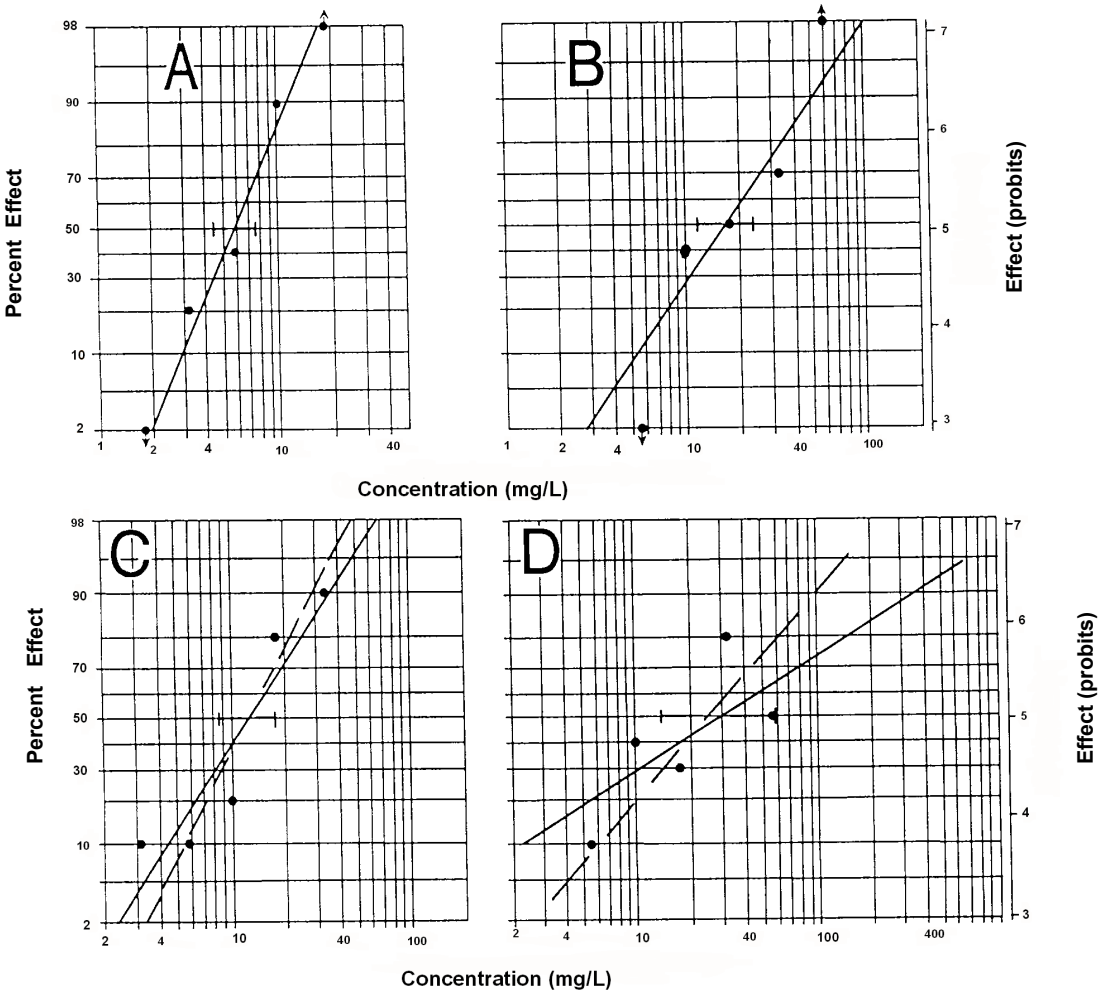


Figure 5 Fitting probit lines by eye, to representative sets of data. The panels A to D show the same data as in examples A to D of Table 2. The dashed lines in C and D would be reasonable fits, but the solid lines would be preferred, and approximate those calculated by probit regression. The 95% confidence limits are shown as a horizontal bar, as calculated by probit regression. For further discussion see text, particularly for choice of lines in some panels.

carries maximum weight. Nevertheless, this potential line only slightly underestimates the EC50 as determined by computerized probit regression. That calculated line has a lower slope, partly to accommodate the greater influence of the three central points. The lower slope is also indicative of greater variation, with a high chi-square of 5.5.

A general message from these examples is that similar estimates of the EC50 are often obtained by making statistical calculations and by using an eye-fitted line. Another apparent conclusion is that a probit line for variable data, properly estimated, might have a lower slope than that derived from fitting by eye.

4.2.4 Effects Among Control Organisms

Key Guidance

- *Most Environment Canada methods for quantal tests allow control effects ≤10%, although certain tests allow up to 30% for particular species. No correction is applied for a control effect within the allowable limit, but higher effects render the test invalid. Cause(s) should be investigated, and the test repeated if possible.*

- *For the special case of Environment Canada's quantal sublethal test with salmonid eggs, a correction is made by Abbott's formula, for eggs which were unfertilized at the start of the test. That correction is satisfactory because fertilization occurs before the toxicant is added. A somewhat similar correction is made in the test for echinoid fertilization.*
- *Available commercial computer programs might not follow Environment Canada's approach for control effects, so the investigator must understand how a program works.*
- *For research or other tests outside of Environment Canada's methods, the best way to deal with a control effect in quantal data is to analyze using a computer program which makes maximum likelihood estimates of the control mortality parameter. Failing that, a correction for control effect might be made with Abbott's formula, but the procedure has basic conceptual problems, from both the biological and the statistical points of view. For the unusual case in which a control effect is greater than the effect in a given concentration, Abbott's formula gives a peculiar answer, and any correction should be to zero percent.*

An occasional 10% effect could occur among control organisms, even under favourable conditions. That would not invalidate tests and no correction should be applied for an effect of that magnitude. Some quantal methods published by Environment Canada specify that a test is invalid if the control shows greater than 10% effect; this applies for rainbow trout, *Daphnia* (EC, 1990c; 1990d), and several other methods. For other test methods using organisms that are more difficult to hold in the laboratory, there can be higher mortality under apparently good conditions. Environment Canada allows 20% control mortality for general-purpose tests with larval fathead minnows and up to 30% for reference tests with certain amphipods (EC, 1992b; 1998b).

For the acute quantal tests of Environment Canada, the usual methods of statistical analysis do not provide any option to correct for control effect (e.g., EC, 1990a,b,c). (A maximum likelihood estimate could allow for the control effect, but is seldom used routinely at present.) With the usual analyses, a test would simply become invalid if the control effect exceeded the limit specified in the instructions. Results would be rejected, and the test could be repeated if desired (and if feasible).

Even if an observed control effect is acceptable according to the EC method, there can be a suspicion that something is wrong with the test conditions or the health of the organisms. A search should be made for any apparent cause, and if found, an attempt should be made to eliminate it. Any laboratory that consistently experienced elevated control effects would, of course, intensify remedial efforts.

Sublethal test with salmonid eggs. This salmonid test (EC, 1998a) is a special case for correction of control effects. There can be high proportions of unfertilized eggs during the initial preparations for the test, but an investigator cannot identify those eggs until later. That failure in fertilization cannot have an interaction with the toxicant, however, because the toxicant is added *after* the procedures for fertilization have been completed. There is no reason (except the toxicant) to expect that eggs, once fertilized, will not develop in a normal manner and proportion. In other words, there can be no physiological interaction between success of initial fertilization and action of the toxicant. In this special case, a correction can be applied for unfertilized eggs, using Abbott's formula which is described in the following text. Some of the major conceptual problems with Abbott's correction do not apply in these circumstances. Therefore, use of Abbott's formula is recommended by Environment Canada for this salmonid test, for any reasonable proportion of unfertilized eggs in the control, including low values of 10% and less. After correction, differences between the control and the experimental concentrations, in the proportions of eggs which fail to develop, are then credited to action of the test material.

In the test of echinoid fertilization (EC, 1992f), an equivalent to Abbott's correction is used in the

analysis to determine the ICp, a procedure that is customary with this toxicity test.

Computer programs. Available programs do not necessarily follow Environment Canada approaches for control effect. Some programs might use sophisticated maximum likelihood procedures to estimate the "true" effect of the toxicant without the control effect (Section 4.5.5). The control effect would still have to be within specified limits, if test results were to be used under the aegis of Environment Canada. Other computer programs might automatically apply Abbott's formula, which would not be appropriate for most methods published by Environment Canada.

Accordingly, an investigator *must* understand exactly how a selected computer program deals with control effects. (Computer programs are discussed in following sections.) The Stephan program (Stephan *et al.*, 1978) and some of its adaptations will not accept any control effects. The programs TOXSTAT 3.5 and CETIS (see reference list under those names) can be directed by the investigator to correct, or not, for control effect. TOXCALC 5.0 applies Abbott's formula in probit regression, where the program considers it appropriate. Choosing a suitable program is the best way to avoid having an undesired control correction applied by the computer. (In any case the control effect would have to be within the limits specified by the test method of Environment Canada).

Use of maximum likelihood estimation. The best way of dealing with control effects is with a computer software package which uses maximum likelihood estimation (MLE, see Section 4.5.5). Programs which offer MLE estimate two parameters to describe the adopted model, and a third parameter for control effect. The endpoint, such as EC50, is estimated for the effect of the toxicant *only*, i.e., without the control effect.

MLE has been available for a long time in major software packages such as SAS (1988; 2000). Such major statistical packages might not be available in all laboratories. The usual software for toxicity tests (at the time of writing, including CETIS, TOXCALC and TOXSTAT) are based on the classic "iteratively reweighted least squares".

Even a sophisticated procedure (true maximum likelihood carried out with SAS) operates only within the confines of the particular test. Although the model separates off the control effect, *it does not compensate for an overall change in resistance of the test organisms*, if such a change were induced by sickness or some similar factor. In plainer words, the EC50 might be representative of weakened organisms with low resistance to toxicants. At present, there is no simple model or modelling procedure that captures the interaction between the effect of the background factor and the effect of the toxicant [28] (see following discussion of Abbott's formula). The remedy is to test under good conditions with a healthy stock of organisms.

Limitations of Abbott's formula. This procedure (Tattersfield and Morris, 1924; Abbott, 1925) is a simple mathematical method of correcting for control effects. Some examples of corrections are shown in Table 1 and Figure 6, while the formula is given by Equation 2. Note that in this formula, proportions are used, e.g., 3 organisms out of 10 would be entered as 0.3.

$$P = \frac{P^* - C}{1 - C} \qquad \text{[Equation 2]}$$

where:

P = the corrected proportion of organisms showing the effect

P^* = the observed proportion of organisms showing the effect

C = the proportion of control organisms showing the effect

Abbott's formula is based on the unlikely assumption that the effect seen in the control is completely separate from the effect of the toxicant, and does not influence it. Evidence has indicated that the assumption is invalid (reviewed in Hewlett and Plackett, 1979), so that use of Abbott's formula

[28] Dr. W. Slob (2003, personal communication, National Institute of Public Health and Environment, The Netherlands) reports that such a method is included in software named PROAST, which is being developed for use by other investigators.

Table 1 Examples of corrections by Abbott's formula, for various control effects in a quantal toxicity test. The hypothetical results are similar to Example B of Table 2, but with less extreme values at low and high concentrations. In the four right-hand columns, the control effect has been changed from zero to 10, 20, and 30% effect. The results at each test concentration have been adjusted for those control effects by Abbott's formula. Probit regression (Stephan *et al.*, 1978) was then applied, to calculate the results in the lower four rows, which are plotted in Figure 6.

Concentration, amount/litre	Number of organisms tested	Number of organisms affected, corrected by hand for a control effect of:			
		zero	10%	20%	30%
56	10	8	7.78	7.50	7.14
32	10	7	6.67	6.25	5.71
18	10	5	4.44	3.75	2.86
10	10	4	3.33	2.50	1.43
5.6	10	2	1.11	0	0
EC50		16.5	20.3	25.2	30.1
Confidence limits		7.85, 31.3	12.0, 31.9	16.9, 43.5	20.7, 53.5
Slope		1.65	1.89	2.38	2.53
Chi-square		0.136	0.286	1.26	0.606

introduces a distorted correction. In a situation of high effect in the control, there could be a combination of effect from the action of the toxicant and whatever factor(s) caused the background effect. For example, organisms that were weak from poor nutrition might be less resistant to the toxicant, which would result in a lower estimated EC50 than for organisms which had enjoyed good nutrition [29].

As mentioned previously, no method is known for correcting the potential interactions of the toxicant and whatever was causing a control effect.

The actions of Abbott's formula are examined using the hypothetical data in Table 1. Corrections are shown in the columns of Table 1, for successively greater control effects from zero to 30%. As the control effect becomes greater, there are increasingly greater changes in the corrected results. The estimated EC50 increases by 80%, slope increases by >50%, but confidence limits remain similar in proportion. The chi-square increases, but remains at least six-fold lower than the critical value. These changes are demonstrated in Figure 6,

[29] There are other problems with Abbott's formula. It adjusts the number of organisms that react, but does not adjust the number tested. The control effect is treated as if it were a constant, and the uncertainty associated with it (its variance) is ignored. Failure to incorporate this into the estimate of EC50 causes an underestimate of the variability of the EC50. If probit regression is being used, the assumption of linearity between probits and the logarithm of concentration is no longer valid when there is a control effect. If the correction is applied for several concentrations, that introduces correlation among the concentrations, although they should be independent. If 100% effect is observed at some concentration, Abbott's formula makes no change in that effect, i.e., all of the effect is credited to the toxicant and none to the cause of the control effect.

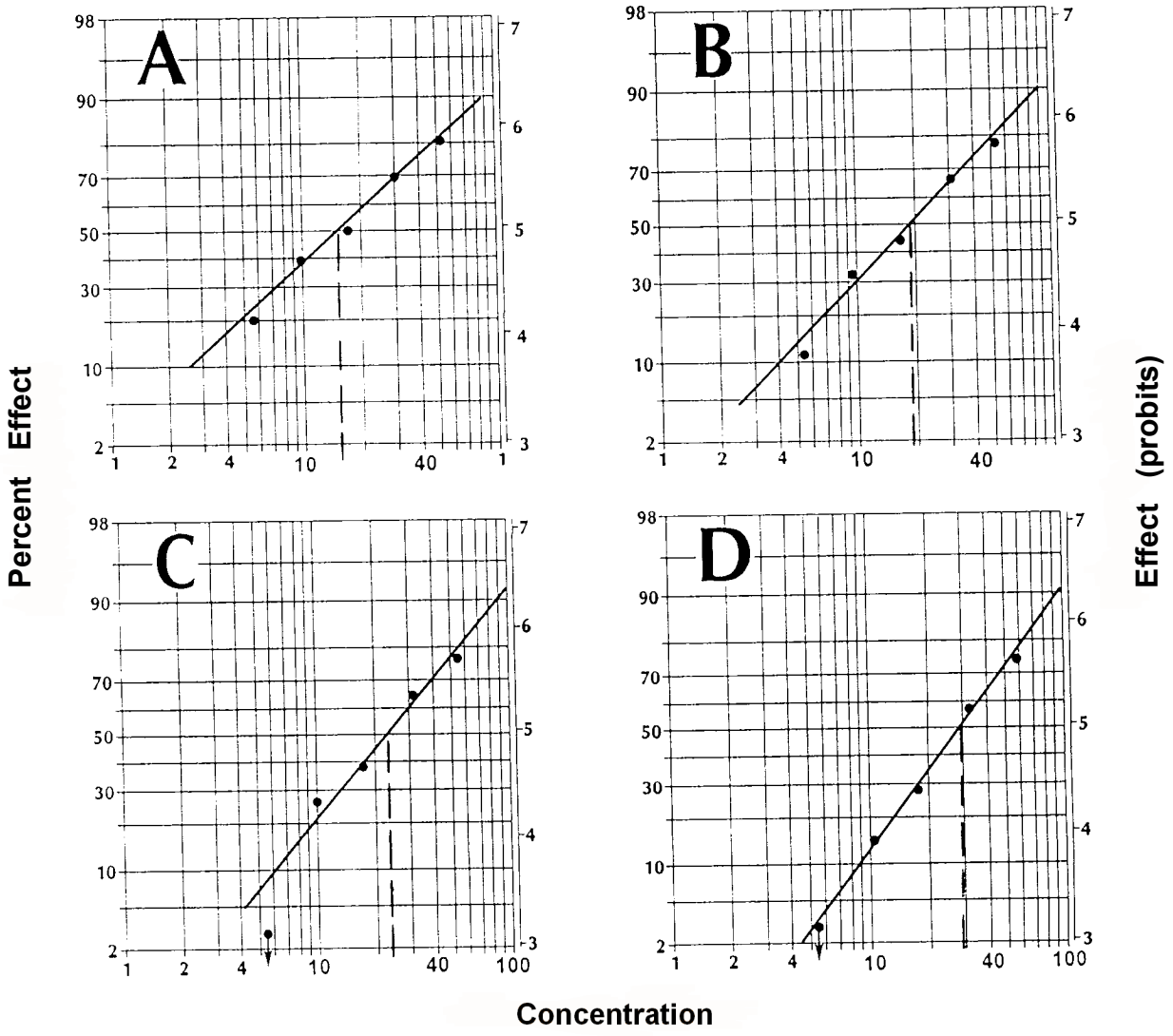


Figure 6 **Results of correcting with Abbott's formula, for control effect in a quantal test.** The lines represent calculated values shown in Table 1. Panel A is an example with no control effect. In successive Panels B, C, and D, the same results are adjusted for control effects of 10%, 20%, and 30%, and the estimated EC50 increases from about 16 to about 30. Slope also increases because with larger corrections, the lower percent effects (low concentrations) are moved downwards towards zero effect, to a proportionally greater extent than are the higher percent effects.

where the probit line moves downward to the right in successive panels, and the slope increases. Other examples might change differently, but EC50 and slope would almost always increase with larger control adjustments.

Abbott's formula with high or anomalous control effects. It should be realized that for organisms amenable to holding in the laboratory, a control effect of 20%, 30%, or more, would cast strong doubt on the validity and usefulness of the ECp. Further, correction for control effect by Abbott's formula would have the major conceptual difficulties outlined here. If those difficulties were accepted by an investigator working for some purpose outside the requirements of Environment Canada, Abbott's correction might be used to correct for control effects up to about 30%. The procedure usually raises the estimated EC50 appreciably, as seen in Figure 6 and Table 1.

For small control effects of ≤10%, it would seldom be desirable to apply a correction, no matter what the purpose of the test. Such a control effect might only represent an accidental, unusual, or random occurrence which had little influence on the EC50 for the toxic material being studied. If that were so, the "correction" would worsen the estimate of the EC50.

If a control effect were greater than the observed effect in a given concentration, Abbott's formula would give a peculiar answer. The observed effect would be corrected to a negative value, unreasonable since it implies that *more unaffected organisms would be present, than were actually tested at that concentration*. Finney (1971) recommends using the corresponding probits for the negative value and continuing the calculations, since this is merely sampling variation. However, investigators might be unable to control the computer program they are using in that way. Some programs have been known to make the correction to a negative value, then ignore the minus sign, use a positive value to create a probit for use in analysis, and carry on with a fallacious calculation of the EC50!

The following recommendation is made here. *If it has been decided to apply corrections for control effect, and if the control effect equals or exceeds an observed effect, and if it is uncertain that the computer program can handle a negative value for effect, then: (a) correct all observed effects by hand; (b) correct the anomalous effect to zero percent rather than a negative value; and (c) enter the corrected effects without the control value.*

An example may be seen in the last column of Table 1. For a control effect of 30% and an effect of 20% at concentration 5.6, Abbott's formula would correct to minus 0.143 or minus 1.43 organisms. Instead, zero was entered.

Hubert (1984) states that "Abbott's formula is only applied to mortality rates which exceed the estimate of the natural mortality rate", but that does not seem reasonable. If an observed effect was less than or equal to the control effect, the observed effect would be left untouched, and the toxicant would be credited with causing it.

It is clear that, with respect to Abbott's formula, an investigator must select an appropriate computer program and understand exactly how it deals with control effects. Among common programs at the time of writing, TOXCALC 5.0 applies Abbott's formula in appropriate situations, while TOXSTAT 3.5 and CETIS provide this as one of several options. The programs of Stephan *et al.* (1978) and OMEE (1995) assume that control effect is zero. If Abbott's formula was being applied, a roundabout but certain way of getting the desired result from a program would be the procedure indicated previously: calculate the corrections for each concentration by hand, then enter corrected versions as if they were the raw observations. No control data would be entered (or zero control effect would be entered if the program required an input for controls). Corrected values would probably involve decimals (e.g., 3.33 earthworms out of 10), but most statistical programs happily accept such fractions.

4.2.5 Confidence Limits on the ECp

Key Guidance

- *The 95% confidence limits of the EC50 must be reported; they estimate internal*

variation of the test. A ratio of 1.3 between the EC50 and confidence limit represents narrow limits and good precision, and ratios of 1.5 to 1.8 are common and acceptable.

- *Confidence limits tell only about the variation in a particular test. They do not indicate overall variation in tests with a given toxicant.*
- *It is good practice to also report the slope of the concentration-effect line, which allows the line to be re-created later if desired. The chi-square for goodness of fit should also be reported.*
- *Confidence limits are narrowest at the EC50, and widen successively at lower or higher degrees of effect.*
- *Because of the variability, there are drawbacks to endpoints with low levels of effect such as EC20. The selected level of effect (p in ECp) should never be within the range expected for control effect, and probably never less than EC10.*

Reported results must always include the 95% confidence limits of the EC50. The only exception would be tests that did not show a partial effect at any concentration. Figure 4 shows that these tests would be analyzed by the binomial method, which does not provide confidence limits. It is also good practice to report the slope of the fitted line relating effect to concentration, and the result for goodness of fit by chi-square. Reporting the slope allows the line to be recreated in the future, if desired; without the slope, there is an inadequate description of the relationship between concentration and effect.

Investigators must always keep in mind that the confidence limits for an individual toxicity test demonstrate only the degree of internal precision of that particular test, with whatever number of organisms was used, under the conditions which prevailed at that time, and with the uncertainties associated with the model. *Those limits must not be mistaken for overall limits of the EC50 for a given test material.* Estimates of EC50 can differ considerably from time to time and place to place, for the same species and similar conditions. For example, if one wished to define the probable limits of toxicity for a particular effluent, the confidence limits from a toxicity test would not tell this. Several samples of effluent would have to be tested. Then, the variation in the endpoints of those tests would be the basis for predicting the limits of effluent toxicity, under the conditions which prevailed over the sampling period. Variation is discussed in Environment Canada's guidance document on the interpretation of environmental data (EC, 1999a).

In Example A of Table 2, most of the calculated confidence limits (upper or lower) differ from the EC50 by a factor of about 1.3 -- good precision in an aquatic toxicity test. In tests with fish, laboratories often find ratios of about 1.3 to 1.5 between confidence limit and EC50, using 10 fish per concentration. Experience indicates that ratios in the vicinity of 1.8 would signify acceptable precision for most purposes [30]. For variable data such as Example D of Table 2, confidence limits might be extreme; some of the upper confidence limits estimated by probit regression are an order of magnitude higher than the EC50. Investigators should be prepared to encounter wide confidence limits occasionally. Sometimes the limits can be improved by choosing a better-fitting model for the data, if they do not conform to the usual pattern. If that is not the case, and the limits are not considered satisfactory, the only other option is to repeat the test.

Sometimes the upper and lower limits might appear to be approximately symmetrical about the EC50 on a logarithmic scale, but some degree of asymmetry would be the normal situation (see following text and Figure 7).

---

[30] Hodson *et al*. (1977) estimate that a typical toxicity test with 10 fish per concentration and three concentrations causing partial effects would have an upper confidence limit that was almost 2.1 times the value of the EC50. Examples A, B, and C in Table 1 have confidence limits that differ from the EC50s by ratios of about 1.3, 1.4, and 1.4. The estimates of variation by Hodson *et al*. (1977) seem somewhat wider than customary findings in many laboratories.

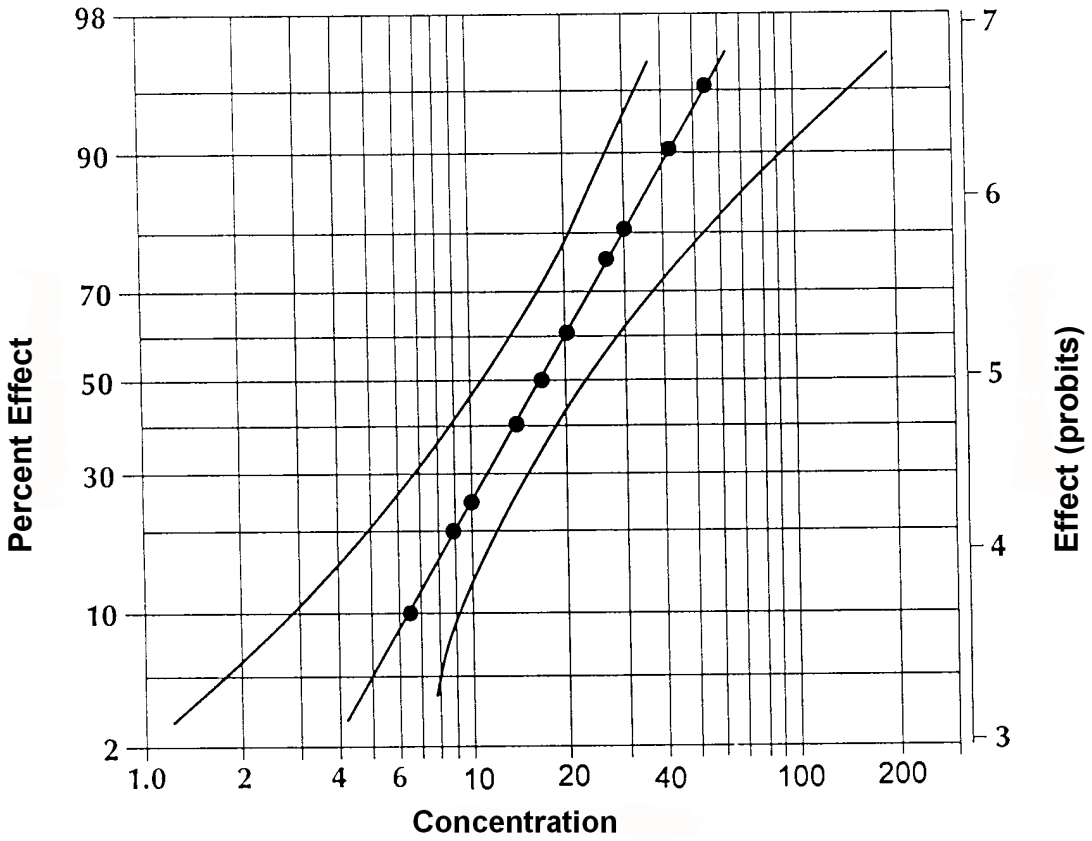


Figure 7 Widening of confidence limits for Effective Concentrations other than the one causing 50% effect. The probit line and confidence limits are derived from Example B of Table 2. Solid points along the probit line show the calculated values of ECp, used to plot the line and confidence limits. The values were estimated by the program TOXCALC.

The span of confidence limits is governed by the slope of the dose-effect line (an indication of variation), by the scatter of observed points about the line, and by the number of organisms at each concentration. If individual organisms were affected at quite different concentrations of toxicant, the probit line would have a low slope, contributing to wide confidence limits. That might happen because of a toxicant's mode of action, without necessarily indicating a procedural flaw. A low slope could, however, be caused by poor procedure such as incomplete acclimation of fish to the dilution water (Calamari *et al.*, 1980).

Precision of the estimated ECp can be improved by increasing the number of test organisms, but major improvements often require more organisms than are practical, as discussed in Section 2.5.

Figure 7 represents results that are regular and the confidence limits are fairly narrow. (Note that the actual data are not plotted in Figure 7. The points shown are the calculated values along the fitted line.) The 95% confidence limits of the EC50 are concentrations of 11.9 and 23.7, differing from the EC50 by factors of about 1.4, considered satisfactory for a toxicity test (see previous text).

Figure 7 shows that the width of 95% confidence limits differs greatly for different percent effects, becoming wider as distance from the EC50 increases. Toward the ends of the concentration-effect relationship, the limits are quite large. This shows why the median effect is a good choice as an endpoint, and why it is not a good approach to adopt endpoints for very low effects, e.g., EC10 which sounds temptingly "protective".

Figure 7 also shows that the confidence limits display some horizontal asymmetry. This represents the usual situation. Limits are originally calculated in terms of the *observed effects* at fixed concentrations, and so at any point in the probit line they are vertically symmetrical about the line (See discussion in Section 9.4). Inverted estimates then produce the confidence limits in terms of *concentration*, as desired by the investigator. Those limits will always be at least slightly asymmetrical, often to a noticeable degree. Limits at the ends of the distribution are strongly asymmetrical.

4.2.6 EC20 or Other Non-median Endpoints

In a quantal toxicity test, it is customary to estimate the median effect (EC50), because that endpoint represents the median or "typical" organism, and is associated with the narrowest confidence limits, i.e., greatest precision. At the same time, there is often a demand for endpoints which are seen to be "more protective", i.e., associated with lower proportional effects, such as the EC20 or EC25. One way to deal with these opposing demands is to accept the median endpoint with its greater precision, then apply an appropriate factor to obtain a concentration that would apply to a smaller fraction of the population of organisms. That procedure has both good and bad features. The more direct approach, using the same general procedures as for EC50, is to estimate the EC20 (or whatever ECx is desired) directly, and tolerate the wider confidence limits.

One caution that should be heeded is not to attempt estimates for a very low value of "p". While an EC01 might sound appealing as a concentration having negligible effect, there are major conceptual difficulties and the variability of the estimate makes it very undependable (Figure 7). There would be questionable validity and meaning, for any attempt to estimate an ECp which was similar to potential control effects. A rule which would seem reasonable, is: *never attempt to estimate an endpoint within the acceptable range of effect in the control(s).* Beyond that, any value for "p" would be suspect if it was below the lowest effect observed for the test concentrations. Thus, the lowest acceptable value for "p" would depend on the data from a particular experiment. It could be less than 10% for a very large experiment, or it might be 20% or even higher in another experiment.

Future progress in estimating ECx for low values of x, is of considerable interest in estimating "safe" or "no-effect" levels of contaminants for humans as well as natural systems. Noppert *et al.* (1994) studied this in response to interest by the OECD, and concluded that the best approach would be modelling of ECx, rather than a hypothesis testing technique. However, they ended by suggesting 5 or 10% as the preferred value of "x", rather than a value closer to zero. Regression techniques to estimate low values of ECx were also concluded to be the superior approach, by Moore and Caux (1997).

A particularly meaningless exercise would be an attempt to estimate a concentration that would just fail to affect any organism (the EC00). That cannot be estimated explicitly because it would depend on population size (one out of a hundred organisms? a thousand? a million?). Nor are statistical procedures designed to provide such an endpoint. (However, Sections 5.2 and 5.3 refer to more sophisticated modelling techniques which extrapolate from acute tests to thresholds of chronic effects.)

Restrictions on type of data suitable for "non-medial" ECps would be those listed at the beginning of Sections 4.2 and 4.2.1. The appropriate value of "p" would be substituted, for example, analysis might require one effect equal to or greater than 20% instead of $\geq$50%.

Several current computer programs provide estimates of non-medial ECps using probit or logit regression. The large computer package SAS does so, and SPSS prints a selection of ECps over the entire useful range. CETIS, TOXCALC, and TOXSTAT do the same, or can be requested to do so. (These statistical packages are found in the reference list under their names.) For the Spearman-Kärber method in these packages, only the EC50 is estimated. The program of Stephan *et al.* (1978) and its adaptations (OMEE, 1995) are also limited to estimating the EC50.

The other approach for estimating low endpoints, would be to start from the estimate of the median endpoint with its greater precision (as shown in Section 4.2.5). Then a factor could be applied, to reach a concentration expected to cause some low partial effect of interest, perhaps a concentration

that was poorly defined in the results of a given toxicity test. For example, a factor could be applied to the EC50, to reach an expected EC20, or even an EC5. The factor could be chosen, based on usual slopes of the probit/logit lines obtained in such tests. (This approach has, in fact, been used for decades to extrapolate from median lethal concentrations to supposed "safe" levels which have been used as water quality objectives. These are the "application factors" described in EC, 1999a.) Use of such factors has the good feature of a starting point which is relatively well defined. There is also the less desirable feature of being mildly or strongly hypothetical, depending on the degree of extrapolation.

4.3 Choice of Methods

Key Guidance

- *Probit or logit regression by maximum likelihood regression (MLE) is the preferred standard method for quantal effects at three or more concentrations, including two concentrations with partial effects. The second choice is the commonly used method of iterative probit (or logit) regression, which provides estimates comparable to MLE. Probit/logit regression is currently recommended for routine use because of availability and convenience of methods.*
- *Some tests might produce only one partial effect, unsuitable for probit/logit regression. The Spearman-Kärber method is recommended for those sets of data. This method should be run with no trimming of data and also with minimum trim (limit of 35% trim).*
- *If successive concentrations produce zero and 100% effects with no partial effect, the approximate EC50 should be estimated by the binomial method. This method should also be used if anomalous results are obtained with the Spearman-Kärber method. The binomial method does not provide 95% confidence limits, but instead estimates conservative limits within which the EC50 would lie.*
- *The moving average method is valid but has the same data requirements as probit or logit regression, which are recommended instead.*
- *For analysis, a variety of commercial and government computer software is available. An operator must fully understand the procedures used by whatever software is chosen. Some software has deficiencies for the purposes of Environment Canada, or needs input of irrelevant material designed for foreign regulatory agencies.*

The following methods of analysis are recommended for tests carried out for programs developed by Environment Canada. The most desirable methods (1) and (2) will not be suitable for most data from routine testing, because they require two partial effects. Secondary methods of analysis are included in the list, for other types of data. The various acceptable methods are described in further detail in Section 4.5.

1. *Probit or logit regression by maximum likelihood* (Section 4.5.3). This is known to be available in the statistical software package SAS (1996). It has the advantage of an unbiased method of allowing for control effect, and estimating an endpoint based only on the effect of the toxicant. Calculation requires two partial effects in the data.

2. *Probit or logit regression by iteration.* The programs use iteratively reweighted regression to arrive at a definitive estimate. Most available computer programs follow this "classical" iterative technique. It provides satisfactory analysis, reaching a solution equivalent to maximum likelihood estimation. The method requires two partial effects.

3. *Spearman-Kärber*. This is recommended only if results cannot be analyzed by the previous two methods. The data must have one partial effect, plus 0% and 100% effects or values near those extremes. Analysis should be run (a) without trimming, and (b) with "automatic" or "minimal" trim ≤35%. The most reasonable of the two estimated endpoints should be selected by inspection of raw results and their plot. If neither endpoint is reasonable, the binomial method should be used.

4. *Binomial.* This is for situations with no partial effect, but 0% and 100% effects. This method would also be adopted for other situations when methods (1) to (3) could not be used. For example, it would be used if there was one partial effect, but the Spearman-Kärber method produced anomalous results, because of lack of 0% and/or 100% effects, or other reasons.

5. *Moving average*. The available program for this requires two partial effects. It might be useful for unusual situations in which probit/logit analysis failed. It would seem to have no particular advantage for other situations.

6. *Litchfield-Wilcoxon graphic method*. Not recommended for definitive reports. Useful for checking computer estimates, for field work, or for training purposes.

True **maximum likelihood estimation** using probits or logits (MLE, item 1) is the most desirable method of estimating the EC50. This method assumes that at each concentration, some proportion of the tested organisms will be affected. It further assumes that those proportions are related in a cumulative distribution function, increasing from 0% effect at low concentrations to 100% effect at high concentrations. The MLE attempts to estimate the values of the parameters in the relationship, that would result in the highest likelihood of observing the data actually collected (see Section 4.5.5). Once defined, the mathematical relationship *predicts* the concentration expected to produce a given effect. MLE can be carried out by the large statistical package SAS, which might not be available in some laboratories, or might not be easily used by investigators.

Iterative probit regression (item 2) is available in authoritative major software libraries, notably SPSS and SYSTAT (programs are listed in the References under their names), and in most other commercial toxicology packages. Because of universal availability, probit or logit regression by iteration is designated here as the customary method for routine use at the present time. The details of the approaches and the choice of logits or probits are considered further in the following text (Sections 4.5.1 to 4.5.6).

The well-known "Stephan program" (Stephan *et al.*, 1978) includes probit regression (item 2), moving average (item 5), and is the only convenient source for the binomial method (item 4). It was developed by Dr. Charles E. Stephan and colleagues of the USEPA in Duluth, Minn., and has been in use for more than two decades. The Stephan program is recommended in many EC method documents, has been generally used in Canadian laboratories, and has been available from workers in those laboratories. The Stephan program has been adapted in various forms. An adaptation at the Etobicoke laboratory of the Ontario Ministry of Environment and Energy, written by Dr. Gary F. Westlake, operates in an early Windows format (OMEE, 1995), has probit, Spearman-Kärber, and moving average methods, and shows a plot of the results (hereafter called the *OMEE program*).

Various commercial computer programs developed in the USA include CETIS and the older programs TOXSTAT 3.5 and TOXCALC 5.0. They can analyze quantal data by various methods, but generally include probit, logit, and Spearman-Kärber. A dependable program in BASIC is described in USEPA (1994a, Appendix I; 1994b, Appendix H; 1995, Appendix H), and is available from the USEPA in Cincinnati, Ohio, or at the web-site http://www.epa.gov/nerleerd/stat2.htm. Some other programs might not be suitable because they contain alternatives which are not appropriate in Canada [31].

---

[31] Design of US commercial computer programs can be influenced by decisions of the USEPA, but might not conform to the practices of Environment Canada. The

Simple linear regression. This type of regression has major limitations and is not recommended. It might seem an obvious mathematical method of fitting a line to quantal data, such as those shown in Figure 5, but it is not valid. The reason is the difference in value ("weight") of the points, which is inversely related to the variation, which in turn increases towards the upper and lower ends of the line. The weights must be incorporated into the fitting process, but there is a "Catch 22"; the weights can only be derived from the fitted line, not from the raw observed effects (see footnote 32). That explains why simple regression cannot be used, and why it is necessary to adopt procedures such as iteration. Occasionally, naive practitioners improperly use simple regression in an attempt to estimate EC50s.

4.4 Comparison of Estimates by Various Methods

Key Guidance

- *Most of the common methods for statistical analysis of quantal tests are likely to produce similar estimates of an EC50 and its confidence limits, for reasonable data.*
- *Examples of hypothetical "good" data were analyzed by various methods. Similar results were obtained by probit, logit, Spearman-Kärber, moving average, and angle methods, and they agreed with those from an eye-fitted line. Estimates of EC50 were somewhat higher by the binomial and Gompertz methods.*
- *Confidence limits were also similar for most of the methods, although the trimmed Spearman-Kärber method showed wider limits. The binomial method did not give confidence limits, but instead provided a range, within which the confidence limits would lie.*
- *For some examples with only one partial effect, the untrimmed Spearman-Kärber method provided good estimates of the EC50s, while the version with trimming could not provide estimates. The binomial method also provided good estimates of EC50.*
- *For some examples of data, which were erratic or lacked zero and complete effects, the untrimmed Spearman-Kärber method gave very aberrant estimates. The estimates with trimming varied with the type of data -- some were excellent and some were improved but still divergent. The binomial method failed.*

Quantal endpoints provided by various statistical methods are compared in this Section. Some relatively good data-sets are used as examples in Section 4.4.1. Section 4.4.2 does the same for data that lack partial effects, a situation frequently encountered in test programs. The comparisons help to explain the recommended methods given in Section 4.3.

The examples in Tables 2 and 3 could be used to assess other statistical programs that become available to investigators.

4.4.1 Estimates for "Good" Data

The hypothetical sets of data shown in Table 2 and illustrated in Figure 8 can be called "good" because they have two or more partial effects, so they can be analyzed by logit or probit regression. Example A is the one used as an illustration in several test methods published by Environment Canada. The first three examples, A to C, have rather regular data, while Example D is erratic.

Spearman-Kärber method has been a common alternative in US procedure, without the limitations recommended here for Environment Canada (Section 4.5.6). Binomial (and moving average) methods are not offered in recent US programs. Instead, they offer "linear interpolation" using two data-points (Section 4.5.9), and that can be satisfactory and equivalent to the binomial method if there are successive zero and complete effects. Investigators would have to make sure, however, that linear interpolation was done with logarithm of concentration as the default situation. Programs might also require inputs of information that is not relevant in Canada, because they formulate their outputs to meet the reporting requirements of the USEPA.

Most of the current statistical programs show similar estimates of EC50 in Table 2, particularly for the regular data. Those computer estimates also agree with the common-sense graphic estimates, shown in the first line of the table. Figure 8 shows that the graphic and computer estimates are reasonable.

For the five computerized probit programs, Table 2 shows identical EC50s for Examples A, B, and C which have fairly regular data. Confidence limits were also very similar. The SAS program by MLE must be considered the best estimate and the standard of comparison. Even for the irregular data of Example D, EC50s from the five programs were quite close. The Stephan/OMEE and CETIS programs matched the SAS estimates very closely. TOXSTAT and TOXCALC yielded an upper confidence limit that was considerably lower than those of the other methods, for the data of Example D.

A more extensive comparison of programs for probit regression was made by Sebaugh (1998), using 50 sets of data. She adopted the SAS method as standard, and found that the EC50s differed by more than 1% in three cases for TOXCALC, five cases for TOXSTAT, and seven cases for the Stephan program. Most comparisons were satisfactorily close. A probit program widely distributed as "freeware" was compiled by the USEPA (described in USEPA, 1995, Appendix H), and agreed with SAS for all 50 sets of data.

The Spearman-Kärber method *(S-K)* can sometimes give answers that agree closely with those of probit regression. For the "good" data of Examples A and B, agreement with probit regression prevails whether the S-K estimates are derived using the OMEE or TOXSTAT program, and whether there is 10% trimming or no trimming (Table 2).

Untrimmed S-K was unsatisfactory for Example C of Table 2, even though that example represented regular data. The problem was absence of both 0% and 100% effects. Without them, both programs for untrimmed S-K gave aberrant estimates of EC50, and the OMEE program did not produce confidence limits. For this same Example C, when 20% of the ends of the data distribution were trimmed off, TOXSTAT estimated EC50 as 13.4, close to the "correct" value of 12.6. The OMEE program also estimated the same value of 13.4 for the EC50, with any trim of 10%, 20%, 30%, or 35% (not shown in Table 2). That endpoint is fairly reasonable. Thus it appears that trimming can be useful in obtaining a reasonable endpoint with the S-K method. It is usually said that 0% and 100% effects are "required" for the S-K method. This example indicates that the program will run without such values, but only provides a reasonable estimate of the endpoint when trimming allows other extreme values (here, the values 10% and 90%) to substitute for zero and complete effects.

Even more extreme examples, which lack zero and complete effects, can be satisfactorily fitted by trimmed S-K. For example, a set of results was postulated which had results for only three concentrations, and the effects were 20%, 50%, and 80%. Untrimmed S-K produced a nonsensical low EC50, but with 20% trimming, a suitable endpoint and confidence limits were estimated (TOXSTAT, minimal automatic trim, not shown in Table 2). This extreme example also shows that trimming can be a useful procedure with S-K.

Apparently, untrimmed S-K can also fail or give peculiar answers for data that are moderately or strongly erratic. In Example D of Table 2, the untrimmed procedures provided grossly divergent estimates of EC50 (4.29 and 5.05 instead of 26.2, last column of Table 2). Clearly, this differs not only from the "correct" answer by SAS, but also from the common-sense, hand-plotted estimate. In fact, the untrimmed S-K estimates of EC50s were lower than the lowest concentration tested, which only caused 10% observed effect. Both the OMEE and TOXSTAT programs for untrimmed S-K performed poorly with Example D (as well as with Example C), presumably because of the lack of 0% and 100% effects.

Trimming 35% of the irregular data off each end of the distribution in Example D improved the estimate to a reasonable value of 24 (by TOXSTAT, compared to the "correct" value of 26.2). Here again, the trimming partially compensated for the lack of zero and complete effects. The OMEE program continued to give aberrant answers with any trim from 10% to 35% (not shown in Table 2).

Table 2 **Four example sets of acute quantal data for toxicity tests.** See text for explanation of methods used for analysis.

	Concentration, weight/litre	Number of organisms affected (e.g., dead) out of ten			
		Example A	Example B	Example C	Example D
	56	--	10	--	5
	32	--	7	9	8
	18	10	5	8	3
	10	9	4	2	4
	5.6	4	0	1	1
	3.2	2	--	1	--
	1.8	0	--	--	--
	Control	0	0	0	0
Graphic estimate	EC50	5.6	17	13	29
Probit, maximum likelihood (SAS)	EC50 (conf. limits)	5.58 (4.26–7.40)	16.9 (11.8–23.7)	12.6 (9.02–18.7)	26.2 (13.1–179)
Probit (Stephan and OMEE) The critical value of chi-square for p = 0.05 and 3 d. of f. is given after the calculated chi-square.	EC50 (confid. limits)	5.58 (4.24–7.37)	16.9 (11.9–23.7)	12.6 (8.98–18.6)	26.6 (13.2–187)
	slope of line	4.71	3.17	3.07	1.32
	chi-sq. (crit. value)	1.11 (7.82)	3.56 (7.82)	3.47 (7.82)	5.52 (7.82)
Probit (CETIS 1.018)	EC50 (conf. limits)	5.58 (4.24–7.37)	16.9 (11.9–23.7)	12.6 (8.98–18.5)	26.6 (13.2–190)
Probit (TOXSTAT 3.5)	EC50 (conf. limits)	5.58 (4.38–7.12)	16.9 (12.4–22.9)	12.6 (9.13–17.4)	26.6 (13.4–53.0)
Probit (TOXCALC 5.0)	EC50 (conf. limits)	5.58 (4.24–7.37)	16.9 (11.9–23.7)	12.6 (8.98–18.5)	27.6 (15.9–85.7)
Logit (TOXSTAT 3.5)	EC50 (conf. limits)	5.63 (4.39–7.22)	16.8 (12.1–23.3)	12.8 (9.36–17.6)	26.5 (13.3–53.1)
Spearman-Kärber, zero trim (OMEE)	EC50 (conf. limits)	5.64 (4.38–7.26)	16.8 (12.4 –22.9)	7.98 (no est.)	4.29 (no est.)
Sp.-Kärber, zero trim (TOXSTAT 3.5)	EC50 (conf. limits)	5.64 (4.40–7.23)	16.8 (12.5–22.7)	10.1 (4.8 –21.0)	5.05 (1.39–18.3)
Sp.-Kärber, 10-35% trim (TOXSTAT 3.5)	EC50 [% trim] (conf. limits)	5.73 [10%] (2.55–12.9)	16.7 [10%] (8.30–33.5)	13.4 [20%] (11.3–15.9)	24.0 [35%] (16.1–35.8)
Binomial (Stephan)	interpolated EC50 (limits of range)	6.22 (1.8–10)	18 (5.6–56)	13.4 (5.6–32)	>5.6 (with warning)
Gompertz (Weibull) (CETIS 1.018)	EC50 (conf. limits)	6.11 (4.43–7.80)	18.6 (12.0–25.2)	14.1 (9.58–19.0)	28.6 (11.2–235)
Angle (CETIS 1.018)	EC50 (conf. limits)	5.54 (4.42–7.47)	17.0 (12.8–22.2)	12.1 (8.81–17.7)	26.8 (14.1–153)
Moving average (Stephan/OMEE)	EC50 (confid. limits)	5.58 (4.24–7.33)	17.2 (12.9–22.4)	13.4 (9.0–24.2)	17.8 (11.9–37.1)

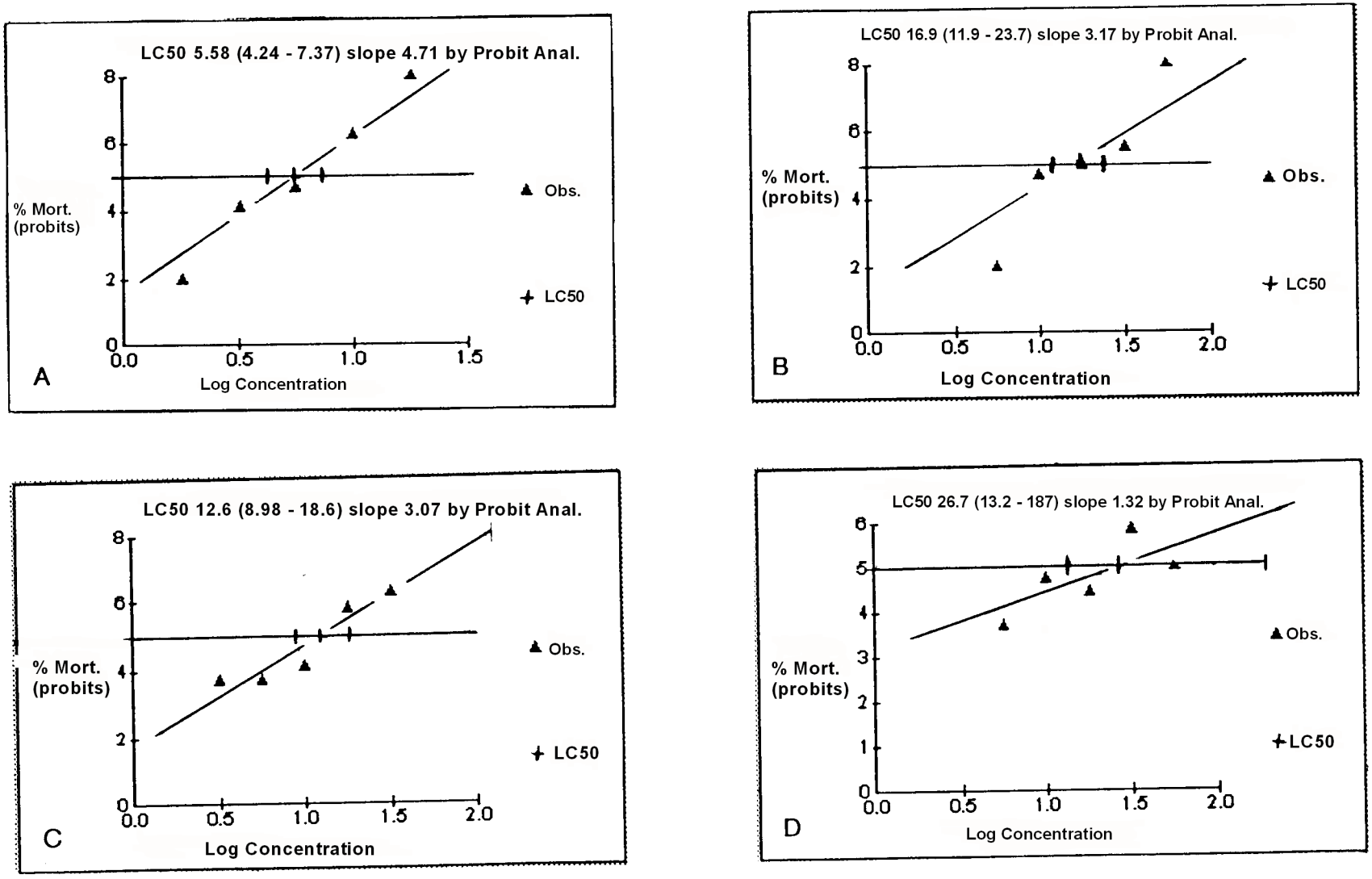


Figure 8 **Appearance of plotted probit regressions for examples A to D from Table 2.** The plots were printed by the computer program of OMEE (1995), and the following items were added: horizontal line at probit 5, the 95% confidence limits, and a fitted probit line.

From these examples, it appears that when the S-K program is being used, estimates should be made by both the untrimmed and trimmed methods. For the degree of trim, an investigator should choose the option which is variously called "automatic trim", "minimal trim" or "automatically minimize trim level" in commercial computer programs (TOXSTAT, CETIS). The programs select the appropriate degree of trim. Results should be evaluated by inspecting the raw data and the plot of those data, then by comparison, choosing the more reasonable of the trimmed and untrimmed S-K estimates. This requirement for the investigator to make a subjective judgment is not a desirable procedure, but it appears necessary for the S-K programs, which do not include any test for validity of the estimated endpoint.

The irregularities with the S-K method are not crucial with the "good" data of Table 2, because S-K would not be used for such results, under the methods published by Environment Canada. All four examples would normally be analyzed by logit/probit methods. The preceding exercise was done to evaluate S-K methods.

The binomial method was also used in Table 2 for illustrative purposes only, because all these examples could be analyzed by probit or logit regression. The estimates by the binomial method were 6 to 11% higher than those by the SAS probit method, for Examples A, B, and C. Of course, the approximate limits of the estimate differ appreciably from the confidence limits of the probit method. For the irregular data of Example D, the binomial approximation failed. The program merely stated that the EC50 would be higher than the lowest concentration tested. It issued a warning: "Obtaining an approximate LC50 by interpolation between two concentrations does not appear reasonable with this [*sic*] data".

Analyses based on the Gompertz and angle transformations are shown in Table 2 although these methods are seldom used. The Gompertz EC50s are noticeably higher than those obtained by other methods, and higher than the common-sense graphic estimate in Examples A, B, and C. The Gompertz model is more appropriate than normal and logistic transformation, if the distribution of effects is asymmetric. Gompertz analysis is analogous to using the Weibull model, which is sometimes found to give the best fit to survival data (Newman, 1995, p. 125). The Weibull model also assumes an asymmetric distribution. Christensen (1984) found that use of a Weibull transformation "generally provides at least as good a fit to experimental data as the probit model", but that is not evident in the endpoints listed in Table 2.

The angle transformation provided estimates that were very similar to the results from SAS and other probit methods. The angle method would appear valid from that evidence, but would not be needed if a good probit or logit method were available. (Angle or angular refers to the arcsine transformation.)

The moving average programs of Stephan *et al.* (1978) and OMEE (1995) provided identical estimates that were also the same or almost the same as probit estimates, for the "good" data of examples A to C. However, moving average gave a rather aberrant EC50 and confidence limits for the irregular data of Example D. As stated previously, the method would not seem to be needed under normal circumstances because the available program has the same requirements for type of data as do the probit and logit methods.

4.4.2 Estimates for Data with Few Partial Effects

More often than not, laboratories encounter test results that have only one partial effect, or none. The results cannot be analyzed by probit or logit regression. The usefulness of other methods is assessed with the examples in Table 3.

The data in Table 3 were developed from those shown in Table 2, by reducing most examples to one partial effect. The two values at the highest concentrations were set at 100% effect, and the two values at the lowest concentrations were set at 0% effect. The only exception was Example D, in which the irregular value of 50% effect at the high concentration continued. The methods shown in the left column were used to analyze these data, or to attempt analysis. As recommended in Section 4.2, the analyses used only one of two successive effects of 0%, or 100%, the one closest to the centre.

Examples A, B, and C cannot be analyzed by probit or logit regression. Nor can the moving average method provide an answer, confirming that this is

not much assistance to investigators as a backup method.

Probit and logit analyses ran satisfactorily for the contorted Example D. All five probit programs provided the same reasonable EC50, and the logit estimate was close. Confidence limits varied somewhat; those from the Stephan/OMEE method extended from zero to infinity, not a very useful estimate.

For Examples A, B, and C, both the binomial and untrimmed Spearman-Kärber methods provided estimates that seemed reasonable, agreeing fairly closely with the hand-plotted graphic estimate. This supports the recent practice of Environment Canada, to use S-K when there is only one partial effect so that probit/logit cannot be used (EC, 2001a; 2004a). It should be noted that the successful S-K analyses were for data that contained both 0% and 100% effects. Trimmed S-K failed in each of these three examples (TOXSTAT), or gave somewhat divergent estimates (OMEE), presumably because trimming was not appropriate for the scanty numbers of observations.

For Example D, the Spearman-Kärber and binomial methods would not be needed, since the preferred methods of probit or logit regression provided estimates of the endpoint and confidence limits. However, the performance of these secondary methods is worth examining. Neither untrimmed S-K or binomial could handle the distorted data of Example D. The S-K method with no trim produced an EC50 that was hopelessly low compared to the values from probit regression; both TOXSTAT and OMEE produced the same nonsensical EC50. The TOXSTAT estimate of EC50, with trimming, was of the correct magnitude, but a little low. The erratic results of the S-K method for Example D support Environment Canada's recent recommendation to use it only when probit/logit regression will not work because of a single partial effect.

These S-K trials with Example D also indicate that both untrimmed and trimmed analysis should be done, and selection between them should be based on judgement using a comparison of the raw results. In some cases, the estimates from both the trimmed and untrimmed S-K method might be unreasonable, and an investigator might have to impose judgement and reject both of them. There does not appear to be a fixed rule that can be applied to detect acceptable S-K results, nor is there a test of validity in the available software programs, so the judgemental aspect must remain.

For the OMEE method of S-K, levels of trim higher than 10% gave increasingly higher and unreasonable estimates of endpoints for Examples A, B, and D, and erratic results for Example C (not shown in Table 3). It appears that there is a flaw in the S-K program of OMEE, and it is recommended that investigators use the Spearman-Kärber versions available in commercial packages of computer software.

4.5 Examination of Statistical Methods for ECp

Key Guidance

- *Quantal effects are binomially distributed and analysis must use appropriate methods. By custom, probit transformation of quantal effect is commonly used to linearize the relation with log concentration. Logit transformation is mathematically superior and provides similar estimates, although it has been less commonly used by environmental toxicologists in the past.*
- *Maximum likelihood estimates are definitive for probit or logit regression, and have the major advantage of segregating any control effects in an advantageous manner. However, maximum likelihood methods are generally available only in large packages of computer software, and so they are not often used for routine analyses in environmental toxicology.*
- *Classical probit or logit regression proceeds by a series of successively improved fits for a line ("iteration"). An acceptable fit is judged by chi-square.*

Table 3 **Four example sets of quantal data with few partial effects.** See text for explanation of methods used for analysis.

	Concentration, weight/litre	Number of organisms affected (e.g., dead) out of ten			
		Example A	Example B	Example C	Example D
	56	--	10	--	5
	32	--	10	10	10
	18	10	5	10	3
	10	10	0	2	0
	5.6	4	0	0	0
	3.2	0	--	0	--
	1.8	0	--	--	--
	Control	0	0	0	0
Graphic estimate	EC50	6.1	18	12.4	31
Probit (SPSS)	EC50 (conf. limits)	-----	-----	-----	28.4 (17.9–48.2) *
Probit (Stephan/OMEE)	EC50 (confid. limits)	-----	-----	-----	28.4 * (0–infinity)
Probit (CETIS 1.018)	EC50 (conf. limits)	-----	-----	-----	28.8 (no est.)
Probit (TOXSTAT 3.5)	EC50 (conf. limits)	-----	-----	-----	28.4 * (19.4–41.5)
Probit (TOXCALC 5.0)	EC50 (conf. limits)	-----	-----	-----	28.4 * (no est.)
Logit (TOXSTAT 3.5)	EC50 (conf. limits)	-----	-----	-----	27.6 (18.7–40.8)
Spearman-Kärber (zero trim, OMEE, TOXSTAT)	EC50 (conf. limits)	5.96 (4.99–7.11)	17.9 (14.9–21.6)	11.9 (10.3–13.8)	9.11 (5.25–25.5)
Sp.-Kärber, 10-35% trim (TOXSTAT 3.5)	EC50 (conf. limits)	-----	-----	-----	23.2 [30%] (18.1–29.9)
Sp.-Kärber, 10% trim (OMEE)	EC50 (conf. limits)	7.02 (5.61–8.79)	24.1 (19.1–30.4)	15.5 (12.6–19.1)	15.8 (---------)
Binomial (Stephan)	interpolated EC50 (limits of range)	6.03 (3.2–10)	18 (10–32)	12.0 (5.6–18)	>10 (with warning) *
Moving average (Stephan/OMEE)	EC50 (confid. limits)	-----	-----	-----	21.1 * (10.0–35.5)

* For Example D, all probit and logit methods warned of significant heterogeneity; most cautioned that the confidence limits were of questionable validity. The moving average method warned that confidence limits were "probably too close". The binomial method warned that the "interpolation does not seem reasonable".

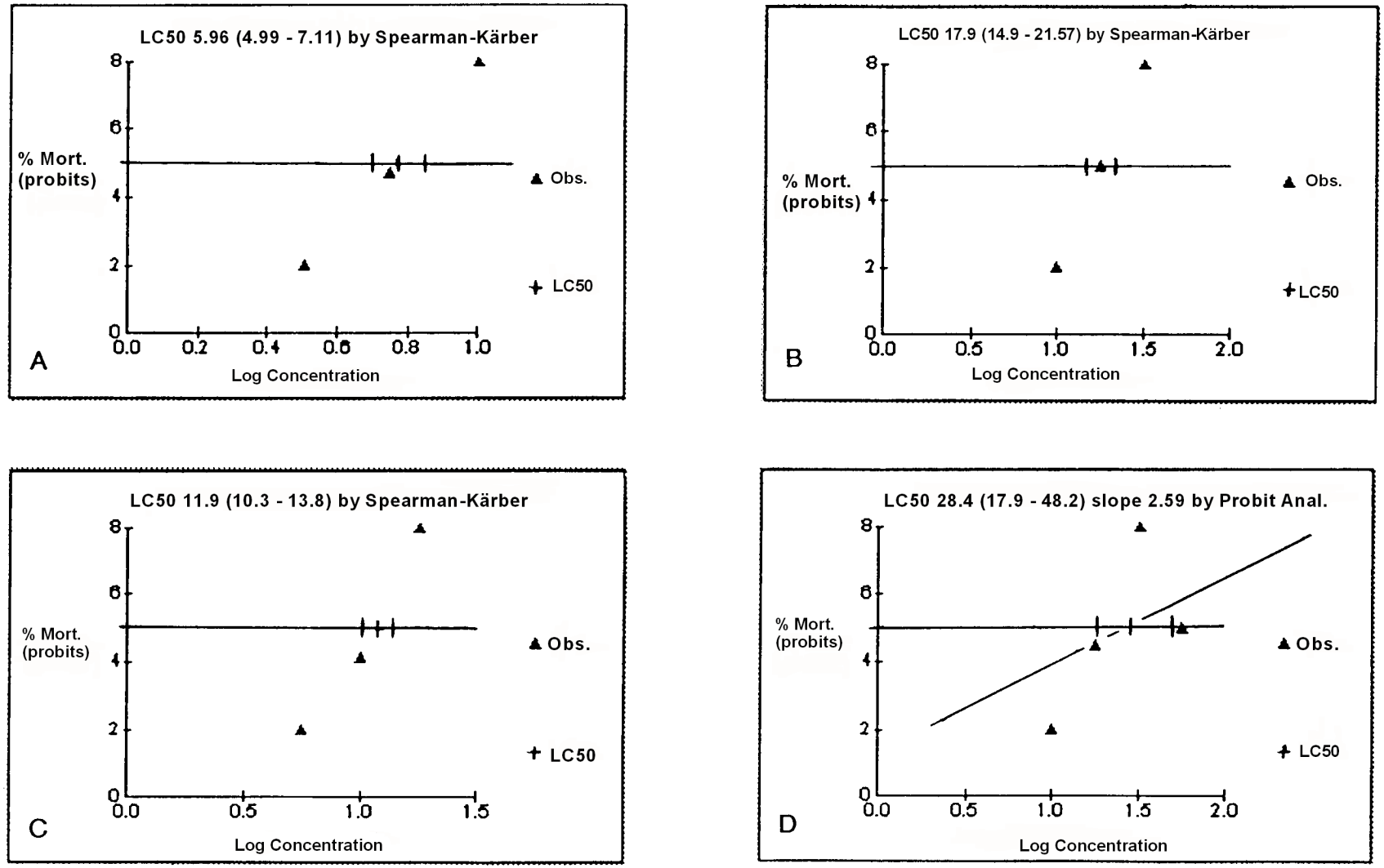


Figure 9 **Plots of quantal data with few partial effects (Table 3).** The graphs were printed by the computer program of OMEE (1995). For A to C, estimates are from untrimmed Spearman-Kärber analysis, and fitted lines are not shown because the method does not use them. Analysis for Example D was by the probit regression program of SPSS.

- *In a limited number of trials, angle transformation of effect also proved to be satisfactory.*
- *The "short-cut" Litchfield-Wilcoxon graphic methods for probit regression are outdated but might be useful for checking computer output or for training new personnel.*
- *The Spearman-Kärber (S-K) method does not estimate endpoints by regression but by weighted averages of midpoints between logarithmic concentrations. The S-K method requires monotonic symmetrical data, and effects of zero and 100%. If the data are not monotonic, analytical programs can impose smoothing. If zero and complete effects are lacking, trimming of data from the ends of the distribution might produce satisfactory estimates from some sets of results. Some recent test methods published by Environment Canada, specify limited use of only the untrimmed S-K method of analysis. It seems desirable to analyze using both no trimming and minimal trimming, then to judge acceptability of each endpoint through comparison with the raw data.*
- *For tests with no partial effects, the binomial method estimates an approximate EC50 as the geometric average of the concentrations causing no effect and complete effect, and takes those concentrations as limits, within which the confidence limits lie.*
- *The moving average method generally performs well but is redundant because the available computer program requires two partial effects, and probit or logit regression can be used instead.*
- *"Linear interpolation" has been designated in the USA as a particular technique. It is essentially the same as the binomial method. Investigators should beware of some older computer programs for this method, which failed to use logarithms of concentrations.*
- *A list of criteria is provided, to evaluate potential new computer programs for analysis of quantal data.*
- *Future analyses could use nonlinear regression if convenient packages are made available for environmental toxicology.*

4.5.1 Probit and Logit Regression in General

Probit or logit regression is a commonly used and satisfactory approach for analysis of quantal data. Logits are superior mathematically as explained in Appendix J, but probits have been commonly used in environmental toxicology. Like all procedures, the method is most effective for reasonably smooth and regular data, and requires two partial effects. The eye-fitted log-probit line (Section 4.2.2 and Figure 5) is a form of probit regression, carried out mentally, without the benefit of calculations.

Explanation is warranted, about moving from a binomial distribution (for quantal data) to analysis based on a normal distribution (as in probit regression).

1. For quantal data such as those from lethal tests, mortality of a single organism is a binary outcome, yes or no.

2. Within a single container, the number of organisms affected *(y)* is the sum of the individual binary outcomes. The variable *y* is a binomial random variable. For that container, the test result is expressed as *y* (the number affected) divided by *n* (the number of organisms in the container).

3. There is usually a series of containers, at different concentrations. If the proportions affected in each container are plotted against log concentration, and the dots are connected, the resulting empirical dose-effect relationship appears to be a cumulative normal distribution (Figure 10, left). It also looks like a cumulative logit distribution function (Figure 10, right), or a Gompertz distribution. This distribution describes the resistance of the sample of organisms to the toxicant.

4. This distribution can now be treated as being normal, or logistic, etc. The binomial effects in the distribution are transformed using probit, logit, or Gompertz, etc. transformations, which address the sigmoidal nature of the dose-effect curve (Figure 10).

5. The resulting *linear* relationship between log concentration and the binomial effect is used to estimate intercepts and slopes. Then the linear model is used in an inverse regression manner (see Section 9.4) to estimate the ECp.

Logistic and probit regression are two of the common methods used for the transformation of step (4); their transformations are indicated in Figure 10 and described further in Appendices H and J. The mathematical formulae for the probit, logit, and Weibull models are shown and explained in OECD (2004).

The left panel of Figure 10 gives a simple illustration of the derivation of probits. The curve is a typical one for percent effect related to log concentration. The horizontal dashed lines represent standard deviations of the cumulated normal curve (half and full standard deviations on the vertical scale of percent effect). At their intercepts with the curve, vertical lines are dropped to a scale that is made uniform in terms of standard deviations. The units are called *normal equivalent deviates (N.E.D. or normits)*. The scale has zero N.E.D. corresponding to the median (50%) effect, and runs upwards and downwards into positive and negative values, as shown at the bottom of the vertical lines. For mathematical convenience in the days of hand calculation, a value of 5 was added to the N.E.D. and the result was named probits, shown at the bottom of the left-hand panel. Now, if the curve is plotted with evenly spaced probits on the vertical axis, it becomes a straight line against log concentration (shown in Appendix H).

Logits are illustrated in the right panel of Figure 10. The pattern and description is similar except that the distribution of effects is assumed to be logistic instead of normal. The horizontal dashed lines are in terms of logits, and that follows through the intercepts to the horizontal scale of logits at the bottom. The result is similar; the curve becomes a straight line when plotted as logits upon log concentration.

After probit (or logit) transformation, statistical analysis proceeds. As described in the follwoing text, the parameters of the probit or logit model must be estimated by rather complex processes, and computer programs are universally used.

4.5.2 Other transformations

There are other models and other transformations. Gompertz and angle transformations can be used in calculations exactly parallel to those for probits. These approaches have the same restrictions as with probits, notably the need for two partial effects.

Results of analysis with the Gompertz transformation were shown in Table 2, and the EC50s were noticeably higher than those obtained by other methods, and generally higher than the common-sense graphic estimates. As mentioned in Section 4.4.1, the Gompertz model and the analogous Weibull model would be more appropriate for data in which the effects were asymmetric. Angle transformation estimated EC50s that were similar to those from probit methods. This transformation could be used but would not appear to be needed if a good probit or logit method was available.

4.5.3 Classical Probit Regression by Computer

In current computer programs for probit regression, the raw (arithmetic) data are entered, and the programs generally carry out the appropriate transformations to logarithm of concentration and probit of effect. Some available programs have idiosyncrasies. With TOXSTAT 3.5, the operator must specify transformation to log of concentration, and must immediately thereafter command the transformation to "run", otherwise the transformation does not remain in force during calculations. TOXCALC has the disadvantage that it does not use probits in its plotted graph.

Most current computer programs for probit regression follow the "classical" methods originally developed for mechanical calculators, before computers were available (Finney 1971; Hubert 1992). Transformation of the proportional effect into probits means that the relationship is linearized against the logarithm of dose, and fitting is simplified to a weighted linear regression. The fit is obtained by successive approximations to the best line *(iteration)*, using a least-squares technique.

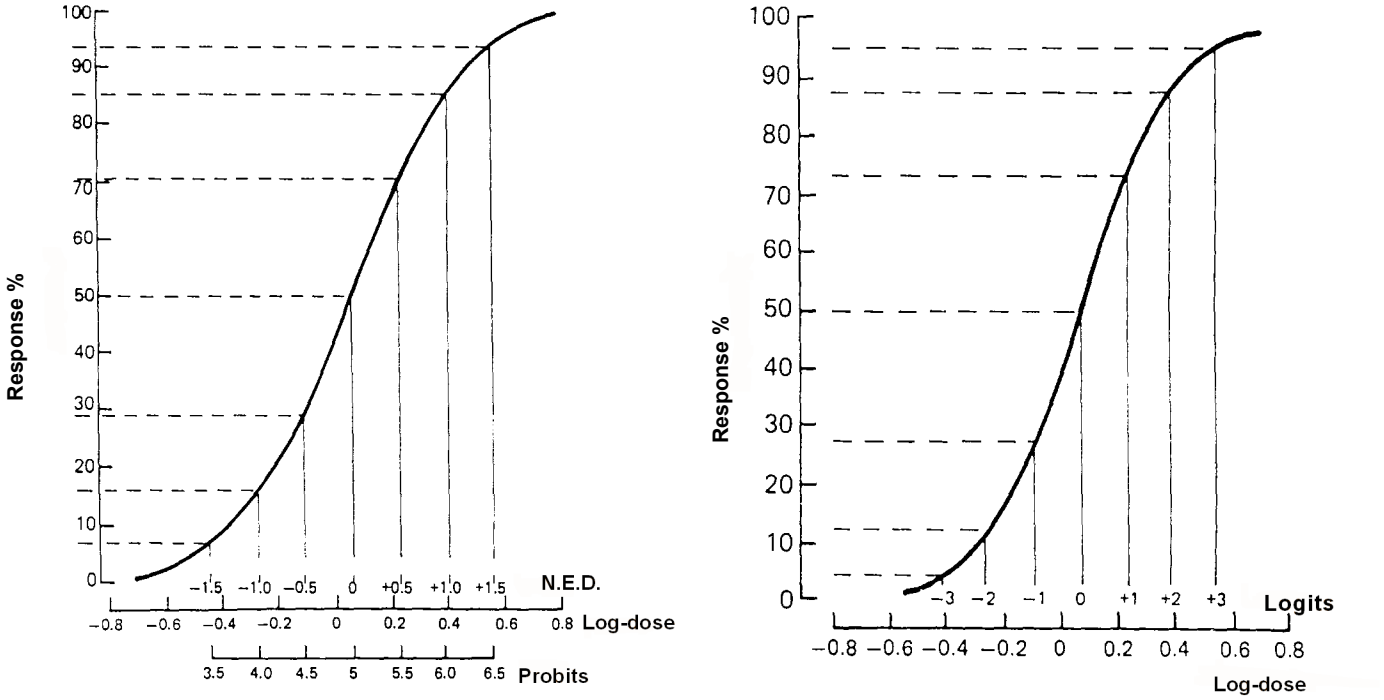


Figure 10 **Graphical illustration of the probit and logit transformations** (from Hewlett and Plackett, 1979)

Computations were feasible with mechanical calculators but remained tedious, time-consuming, and prone to error. The iterations were required because the weights (relative values) for the observations were initially unknown, and depended on the parameters that were yet to be estimated. Numbers of individuals at each concentration contribute to the weighting process. The procedure can be described as "iteratively re-weighted least squares" [32].

[32] These mathematical procedures were designed for the limited capabilities of mechanical calculators. The procedures are fairly complex. (1) The computer does a rough fit of a line to the raw data using logarithms of concentration and probits of effect. (2) It reads the expected probits (= % effect) from the rough line. (3) It "looks up" in a table of constants, initial weights for the observations on the basis of the expected probits, then assigns those weights to the observations. (4) It "looks up" working probits on the basis of expected and observed probits. (5) It fits a better line on the basis of the working probits, weighting factors, and numbers of organisms. This provides the first estimate of EC50, confidence limits and chi-square as a measure of fit. (6) Another cycle of steps (2) to (5) is done, using the working probits of (4) as if they were raw data. In other words, new values are "looked up" for the weights and working probits. (7) Process (6) repeats itself until the answers approach stabilization ("converge"), and the final calculations are adopted.

This procedure of fitting by iterated, reweighted least squares is a way of performing calculations to achieve a maximum likelihood solution. Sometimes the number of cycles is under operator control, otherwise the program has a built-in criterion for stopping the cycles. Sometimes two or three cycles are adequate; for good sets of data, results of successive cycles change little. For irregular data, there might be a "failure to converge" (lack of suitable fit) in 20 cycles; there would be little reason to pursue a fit beyond that. Anomalous data sometimes produce peculiar results after half-a-dozen cycles, such as an unrealistic very low slope and wide limits, as the program attempts to represent the range of results.

Although cumbersome by hand calculation, this classical method can be carried out painlessly with computers. The procedure is not maximum likelihood estimation but it obtains results that are essentially the same; i.e., it "attains a maximum likelihood solution" for the estimate of EC50 and parameters of the line. The classical iterative procedure was once considered the "most efficient" technique for good distributions of data that are log-normal (Gelber *et al.,* 1985).

There is one important weakness of the iterative least-squares technique. It cannot be extended to deal effectively with any effects in the control. That can be done by appropriate models using maximum likelihood techniques (Sections 4.2.3 and 4.5.5).

4.5.4 Assessing Fit with Chi-square

The fit of the probit line is shown by the computed value of chi-square, which must not exceed a critical value if the line and the estimates are to be accepted. Computer programs normally carry out these calculations, but the investigator should make sure that the chi-square value is satisfactory. The assessment by chi-square is approximate, since it would require at least 30 individuals per treatment "to be statistically justified" (Hubert, 1992).

Critical values of chi-square may be found in standard statistical texts. The degrees of freedom in a toxicity test are two less than the number of concentrations tested. The following tabulation could be used for a probability value of 0.05.

Degrees of freedom	Critical value, chi-square
1	3.54
2	5.99
3	7.82
4	9.49
5	11.1
6	12.6

For the four examples in Table 2, the number of concentrations is always five, so degrees of freedom are three. The critical value is 7.82. If a calculated chi-square exceeded that value, the data would be significantly heterogeneous and the line would not be an acceptable fit. All four examples in Table 2 are acceptable.

It is also desirable to make a visual check of the computed probit line. It should be compared with the eye-fitted line, created for this purpose (Section 4.2.2). The OMEE program provides a graph of results, and other programs might do the same. If not, the computed line should be plotted alongside the hand-drawn line, on log-probit paper. Plotting can be done with ease since the slope that is calculated by the program represents the rise in number of probits for a run of one logarithmic cycle of concentration. Starting with the known point of EC50, one log cycle and one probit are scaled off (upwards or downwards or both) to locate a second point for the line (or second and third points). Plotting is even easier if a computer program reports a series of endpoints (EC10, EC20, etc.), as is done by SAS, SPSS, CETIS, TOXCALC, and TOXSTAT.

4.5.5 Maximum Likelihood Estimates

Maximum likelihood estimation *(MLE)* is an objective technique for choosing the values of parameters, for a model that is being used to fit a set of data. The parameters are chosen to maximize (in a selected model) the likelihood of observing the data that were actually collected.

For a quantal toxicity test, the number of organisms affected at a given concentration follows a binomial distribution. The parameters of the binomial distributions are assumed to be related to the concentrations by a function, usually the normal or the logistic. Under those conditions, the maximum likelihood estimates are found to depend on two equations. Today the equations can be directly solved to select the values of parameters, using a personal computer and modern statistical software packages such as SAS [33].

Use of MLE in toxicity tests is only a small part of its general application. Models can be adopted from a variety of patterns to fit different kinds of data,

[33] Finney (1978), the pioneer in the field, welcomed the arrival of modern computing machinery with the statement: "One of the greatest gains to statistics from computers is the ease of initiating and executing iterative calculations. ... Moreover, the classical probit and logit iterative regression calculations can be replaced by direct optimization techniques that lead to the same answers expeditiously and more accurately than before."

while the MLE techniques apply throughout. Thus, MLE could be useful for analyzing various kinds of toxicity tests, whether quantitative or quantal. For example, the model could be a regression of the weight of organisms on the logarithmic exposure concentrations. Alternatively, it could be a distribution function that described the probability distribution of a single set of observations. Here, models for quantal tests are being considered (i.e., probit regression) representing one very useful application of MLE.

For probit regression, MLE is "equivalent" to the older method of iteratively reweighted least squares (Jennrich and Moore, 1975). In other words, MLE reaches estimates for EC50 and confidence limits that are very similar to those of the iterative technique described in Section 4.5.3. MLE is, however, mathematically more elegant, and should be regarded as the definitive approach. In the older iterative probit regression, there are two parameters of interest, the slope and intercept. In modern programs for maximum likelihood, those parameters are replaced by their equivalent parameters, the mean and variance. A likelihood function is manipulated to express the parameters as functions of the data. Calculus is used to set the first derivative equal to zero, then equations are solved for the maximum likelihood estimates of parameters.

Because MLE is a standard technique of statistical analysis, it is included in major statistical software packages. Probit regression using MLE is specifically available in the major statistical package SAS (2000) and perhaps in others. The SPSS and the toxicological programs TOXSTAT, TOXCALC, and CETIS appear to use the older iterative line-fitting. Environmental toxicologists would no doubt find it convenient to have MLE included in software packages that were tailored to their needs.

Control effects. A major advantage of probit or logit regression with MLE is the ability to estimate a control effect as a separate variable, and use only the toxicant-induced effect to estimate the ECp. The observed effects are the sum of two sources of effect, and the level of control effect is included as one of the parameters to be determined in the model. The more complex model has three equations to solve (for mean, variance, and control effect). Two rates of effect are estimated, one of them a baseline or control effect that was not attributable to the toxicant. The other rate is the incremental effect due to the toxicant alone, and it is used to estimate the EC50 without the effect of the baseline condition showing itself in the control.

Maximum likelihood estimation is the best mathematical method of dealing with control effect. However, as pointed out in Section 4.2.4, it cannot remedy any interactions that are biological instead of statistical. For example, disease might cause a control effect, and might also weaken the test organisms' resistance to the toxicant. The analysis would estimate an ECp that was statistically valid, but for weakened organisms.

4.5.6 Spearman-Kärber

The Spearman-Kärber procedure *(S-K)* is recommended here for quantal data which include (a) one partial effect, and (b) 0% and 100% effects. In other words, the method can be used when probit/logit methods will not function because the data do not include two partial effects. This method is available in most commercial programs such as CETIS and TOXSTAT, and at the web site http://www.epa.gov/nerleerd/stat2.htm. It is also available in the OMEE program, although that version of S-K seems to have a procedural flaw in certain cases of irregular data, and is best avoided.

The Spearman-Kärber method was introduced for use in environmental toxicity tests by Hamilton *et al.* (1977), and has a much different mathematical approach than probit regression. The S-K method estimates the EC50 from weighted averages of the *midpoints between concentrations*, on a logarithmic scale. The weight applied to each midpoint is the *change in proportion of effect* between the two concentrations. The concept is similar to estimating the mean of a frequency distribution, whereby the class mid-points are multiplied by the proportion responding within each class. (Further explanation in Appendix K.)

The S-K method can deal with unequal spacing of concentrations on the logarithmic scale, and also with unequal numbers of organisms at the various concentrations. There is no intrinsic method of dealing with an effect in the controls.

Confidence intervals can be estimated if there is at least one partial effect. They are estimated as ± 2 SD of the EC50. This assumes that the estimated

EC50 is distributed as a normal random variable (Miller and Halpern, 1980). The limits are "not likely to be far wrong" unless the number of observations is low (Finney, 1978).

The test has the following requirements or assumptions of monotonic, symmetrical data which include 0% and 100% effects.

- **Monotonic data** is a requirement. If effects decrease from one concentration to a higher one, then the effects are averaged and the result assigned to both concentrations. This smoothing procedure cycles through the set of data until it becomes monotonic (Appendix K). The smoothing does not affect the value calculated for EC50, but it will affect the confidence limits.

- **Symmetry** is an assumption of the method. If the distribution of effect is asymmetric, the S-K method does not estimate a true EC50. Even if trimming were employed, the estimate of EC50 would only be reasonable if the central (untrimmed) part of the distribution were symmetrical.

- **Zero and complete effects** are required, and this is somewhat related to the assumption of symmetry. Without these extreme effects, the untrimmed method will fail, or at best will produce anomalous results. Trimming can sometimes remedy the lack of zero and complete effects, if there are low and high effects such as 10% and 90%.

- **Trimming** is an attempt to correct for non-symmetry in the tails of the dose-effect curve. Trimming can be invoked in order to delete the extreme values and use the central data. This can be useful if there are unexpectedly large proportions of organisms in either tail of the distribution, i.e., many of them reacted at low concentration, or many failed to react at high concentration. Hamilton (1979; 1980) studied these situations and found that a small amount of trimming resulted in a standard error for the estimated EC50 which was much smaller (i.e., more optimistic) than for other reference methods such as maximum likelihood probit or logit analysis. Extensive trimming further decreased the standard error, but raised the estimate of EC50. Hamilton suggested trimming of 10–20% for cases with erratic results in the tails of the distribution, but avoidance of trimming for data with regular distribution.

Recent test methods published by Environment Canada (EC, 2001a; 2004a) are similar to the recommendations of Hamilton (1979; 1980), but more restrictive since they do not allow trimming. Discussion at a meeting of Environment Canada's Statistical Advisory Group had expressed doubts about "fitting a statistical model to make the data look more robust than [they] really [are]". Other statements were: "Don't try to make a silk purse from a sow's ear"; and that trimming led to "difficulties with the variance, hence with confidence limits" (Miller *et al.*, 1993). As mentioned below, prohibition of trimming is probably overly conservative.

In any case, the Spearman-Kärber method is recommended here, only for those quantal tests which produce one partial effect plus zero and 100% effects. For such sets of data, the S-K method is preferred over the binomial method because it calculates confidence limits that can be considered legitimate.

For "good" sets of data, the S-K method can give answers that are very similar to those from probit regression, but might not yield trustworthy answers under some circumstances. Comparisons in Sections 4.4.1 and 4.4.2 showed that this was sometimes true for both trimmed and untrimmed variations of the procedure. The untrimmed method could give very peculiar answers for data that were moderately or strongly erratic (both Examples D in Tables 2 and 3). The trimmed procedure sometimes provided a better estimate of the endpoint, but in some cases of scanty data it failed to provide an estimate (Table 3, Examples A–C).

The untrimmed S-K is almost certain to give an anomalous EC50, perhaps without confidence limits, if the data lack 0% and 100% effects. An analysis could be attempted using minimal trim, if the data-set contained quite low and high values ($\leq$20%, $\geq$80%) as well as a central partial effect. A trim of 20% is likely to produce a reasonable estimate of EC50 and confidence limits.

Apparently, from the examples discussed in Sections 4.4.1 and 4.4.2, the most reasonable way to use the S-K method is to make an estimate by the procedure with no trimming, and another estimate using minimal trimming to the degree selected by the computer program. A choice should be made from the two estimates (if different), by comparing with the raw data and a plot of those raw data. That involves judgement, but it appears to be unavoidable. No test of the validity of the estimate is provided in the computer programs.

The S-K method was named in early methods documents published by Environment Canada, but was not recommended as a method of analysis (e.g., EC, 1992b). However, in EC (2001a), its use was specified for quantal data sets with only one partial effect, which could not be analyzed by probit/logit regression. In the most recent of Environment Canada's test methods (EC, 2004a,b,c) there is new guidance that the S-K method with limited trimming can be used with caution for sets of data with only one partial effect. Investigators should follow the limitations for the S-K method in particular test methods of Environment Canada. A useful approach would include judicious use of trimming as suggested in the preceding paragraph.

An investigator should carefully check the operating procedures used by any program for the Spearman-Kärber method. Those available at the time of writing allow the investigator to choose between no trim and trimming. Here, it is recommended that both those options be used. Some programs have allowed the user to specify the level of trim that will be used (e.g., OMEE). Others (TOXSTAT, CETIS) offer an "automatic" procedure in which the program selects the lowest satisfactory level of trim. That "automatic" or "minimal" option is recommended here.

4.5.7 Binomial Method

The binomial method is a known mathematical procedure, and is currently available as a convenient computer package for quantal analysis, in a program by Stephan *et al.* (1978) and also as modified for a Windows format (OMEE, 1995). It is recommended here for the numerous sets of data in which one concentration results in zero percent effect on the test organisms, and the next higher concentration causes 100% effect. It is also to be used for a set of data which has one partial effect, but cannot be satisfactorily analyzed by the Spearman-Kärber method.

The mathematical procedures are very simple. With no partial effects, the binomial method approximates the EC50 as the mean of the logarithms of the two concentrations causing 0% and 100% effects. It does not estimate confidence limits, but uses those same concentrations as a conservative (wide) range within which the EC50 lies. True confidence limits would likely be well within that range (see below).

The basic calculation of an EC50 can be done easily without a computer program by calculating the average of the logarithms of the two concentrations which bracket the EC50. This is the geometric mean, which can also be estimated by multiplying the arithmetic values for the two concentrations, and taking the square root, as in Equation 3.

$$EC50 = \sqrt{(C_L)\,(C_U)} \qquad [\textit{Equation 3}]$$

where:

C_L = the arithmetic value of the "lower" concentration with no effect

C_U = the arithmetic value of the "upper" concentration causing complete effect

The range within which the EC50 is presumed to lie is given by the same two concentrations.

This binomial method is, in fact, a simple *linear interpolation* on a logarithmic scale of concentration. The name *binomial method* has been retained here to keep the label which has been used for a long time. The name "linear interpolation" has been kept separate (next section) to avoid confusion, because it has been used in the USA to describe a particular technique that does not always represent satisfactory practice.

The binomial method is of great usefulness, as was demonstrated for the data of Table 3, because it is common to have no partial effects when testing industrial effluents. If the concentrations were spaced with reasonable closeness, a test producing such data should not be regarded as deficient, but rather as a legitimate sharp and uniform response by the test organisms. This can indicate a very precise

test, as discussed by Stephan (1977) [34], and in such cases, use of the binomial method is recommended.

In tests with no partial effect, the true confidence limits are usually well within the concentrations causing 0% and 100% effect. Given a finer gradation of concentrations, the lower limit could be as high as the concentration causing 30% effect, and the upper limit could be as low as that causing 70% effect (Doe, 1994) [35]. This was demonstrated in Table 2, in which the limits of the binomial method were much more conservative (wider) than the true confidence limits from the probit method.

The binomial method is also recommended if the data show one partial effect, but cannot be analyzed by the Spearman-Kärber method because 0% or 100% effect is lacking, or other reasons. If the probit or Spearman-Kärber method is valid, the binomial method need not, *and should not*, be used. Nevertheless, the binomial method will function, and approximates the EC50 obtained by more sophisticated calculations. Comparisons in Table 2 showed that binomial EC50s were somewhat higher than those by probit or logit methods.

4.5.8 Litchfield-Wilcoxon Graphic Method

In decades before the 1970s, this graphic "short-cut" method of probit regression was well-used because computers and scientific calculators were not widely available. The method started with a hand-drawn regression, then tested goodness of fit and estimated confidence limits by simplified calculations and nomograms (Litchfield and Wilcoxon, 1949). Appendix L gives further description of the techniques.

The method is not recommended here for definitive estimates, but it can still be useful for checking the estimates of EC50 and confidence limits produced by a computer program. Indeed, the initial hand-drawn line is recommended as the first step in any analysis, to check for reasonable computer estimates (Section 4.2.2).

The Litchfield-Wilcoxon method might also have a useful function in training new personnel. Going through the steps of the graphic method could provide insight on how the endpoints and confidence limits are influenced by various types of data. It could help people to recognize anomalous results from a computer program.

The method was formerly useful for initial estimates under field conditions if there was no access to computer programs. Lap-top computers now remedy that situation.

4.5.9 Linear Interpolation

Although the words "linear interpolation" signify an ordinary and widely used technique, it is listed here as a separate method because the USEPA has designated it as a distinct category with a distinct statistical procedure (USEPA and USACE, 1994) [36]. The method is also sometimes called the "Graphical Method" (USEPA, 2000a). Accepting for the moment, that these names apply to the particular US procedure, it can be said that the procedure has no particular advantage, and is not recommended. The methods recommended for Environment Canada purposes would be probit or logit, Spearman-Kärber, or binomial, depending on the number of partial effects. The US method of linear interpolation is the exact equivalent of a binomial estimate if there are only two concentrations to deal with, one giving an effect below 50% and the other above 50%.

---

[34] Stephan (1977) covers most techniques for estimating quantal endpoints, as a background for his computer program. He justifies the binomial and moving average procedures, and explains why environmental toxicologists should not be overly concerned when they do not obtain two partial effects in a quantal test. Those effects were important in pharmacological work, which gave rise to probit regression, because the investigators needed to assure themselves about slopes of the probit lines, before estimating relative potency of two substances. In the kind of toxicological work discussed here, Stephan points out that useful endpoints can be obtained without any partial effects.

[35] Confidence limits for a test with one partial effect can be read from tables provided by van der Hoeven (1991), but only for special circumstances. The ratio of successive test concentrations must be two. There must be no effect at the concentration immediately below the partial effect, and there must be complete effect at the concentration immediately above the partial effect. A "fairly complicated numerical procedure" would be required in other situations.

[36] Available at http://www.epa.gov/nerleerd/stat2.htm

Linear interpolation (and the binomial method) assume a change in effect that is linear with the logarithm of concentration. For data with no partial effect, either method carries out the equivalent of drawing a straight line on a graph between the logarithms of concentrations causing 0% and 100% effect, then interpolating to the logarithm of concentration causing 50% effect. This is the equivalent of the averaging done in Equation 3.

A further reason for avoiding the "linear interpolation method" is that some computer programs were based on arithmetic values of concentration, and hence gave erroneous estimates. That improved, and logarithms were used by USEPA and USACE (1994).

The "linear interpolation method" is further outlined in Appendix L. The appendix includes a more generalized method of linear interpolation, which would handle data sets with partial effects. It could conceivably be useful in an unusual situation.

4.5.10 Moving Average

This method is not recommended for Environment Canada programs, but elsewhere it has been considered as a possible choice for analysis of quantal data, and Stephan (1977) considered it "the method of choice" for aquatic toxicology. The moving average method estimated EC50 and confidence limits which were identical or similar to those of probit methods for "good" data of Table 2 (Section 4.4.1). However, for irregular data, it gave anomalous estimates compared to other methods (Examples D in Tables 2 and 3).

The method, developed by Thompson (1947), needs results from at least four treatments and they must be at equal geometric/logarithmic intervals. Furthermore, it assumes that there is a symmetrical distribution. It can estimate the EC50, but not any other value of "p" such as EC25.

In theory, the moving average method should estimate the EC50 with one or no partial effects, although in the examples of Section 4.4 it would not provide confidence limits without at least one partial effect. In practice, the available standard program for moving average (Stephan *et al.*, 1978; OMEE, 1995) does not run unless there are two or more partial effects. Probit or logit regression would run on the same data and is recommended here. The moving average approach has some limitations, described by Finney (1978) who comments that "its inherent weaknesses are scarcely balanced by its computational simplicity in an age when computing is so cheap". Possibly, there might be some unusual sets of data for which probit regression would not be suitable, and for which the moving average method could provide an analysis.

For a given set of data, the moving average method estimates several sets of EC50s and confidence limits, one set for each "span" used in calculations. The span indicates how many intervals between concentrations are included in the calculations. The Stephan program prints the results from calculations using several spans, so the investigator can examine the changes produced by various spans. The most appropriate one is indicated by the lowest value of "*g*", which is printed by the Stephan program. The OMEE version of the Stephan program selects the most appropriate span and indicates which one was used. Finney (1978) suggests "the largest possible span" without specifically defining "possible".

4.6 Evaluating New Computer Programs

The data in Tables 2 and 3 might be used to evaluate future new computer programs for estimating ECp. Results could be compared with those shown in the tables, particularly Table 3 for data lacking partial effects. If in doubt about the usefulness of a new program, an investigator could analyze other less-than-perfect sets of data with the new program, and compare with output from one of the more powerful programs such as SAS or SPSS.

The criteria that can be used to evaluate a computer program for analyzing toxicity tests, were listed by Atkinson (1999, slightly reworded), after he reviewed available programs.

- Freedom from non-relevant coding and reporting format (e.g., specifications of USEPA).
- Requirements and cost of equipment and software.
- Requirement for purchasing additional programs (e.g., EXCEL).

- Quality and friendliness of directions.
- Constraints on analyses and data entry.
- Restrictions on number of concentrations and replicates.
- Methods included for desired endpoint calculation (e.g., logistic, Williams' test).
- Methods contrary to recommendations of Environment Canada.
- Suitability of default settings.
- Proper use of logarithmic scale of concentration for calculations.
- Treatment of unequal numbers of replicates.
- Inappropriate control adjustments for quantal tests.
- Existence and usefulness of graphical presentations of data.
- Goodness-of-fit tests included.
- Availability of summary statistics and simple tests.
- Correct confidence limits (compared to other methods).

All of these criteria might not apply to a particular program, but they provide a partial framework for assessment. The last item might be expanded somewhat. Although one assumes that the estimated endpoint will be correct, as well as the confidence limits, the correctness should be tested by comparing results from accepted programs and from examining the plotted data. It was reported in Section 4.2.3 (footnote 27) that one laboratory found the output of a newly purchased computer program to be erroneous, after comparing with hand-drawn graphs (K.G. Doe, personal communication, EC, Moncton, New Brunswick).

4.7 Nonlinear and Other Possible Future Methods

For the immediate future, probit or logit regression seems likely to be the method of choice for estimating ECps in toxicity tests of standard design. However, new approaches such as nonlinear models are developing for analysis of quantal data. To some extent, new quantal methods are coming as "extras" from methods developed for regressions with quantitative data. A relevant example is the adoption of an approach for linear/nonlinear regression by Environment Canada (Section 6.5.8).

Whatever methods develop, they must estimate the EC50 and its confidence limits, if the testing is within monitoring programs of Environment Canada. Good computer programs will also provide a description of the fitted line (such as slope if it is a straight line) and will measure the goodness of fit.

A procedure already available is a complete program for analysis of toxicity data offered by Kooijman and Bedaux (1996). The program is primarily for nonlinear regression but it is said to provide analyses of quantal data on mortality (LC50s), effective concentrations (EC50s), and effective times (ET50s), all with confidence limits (see Section 5.1).

Individual statisticians have used nonlinear models for many years, for analysis of quantal data and determination of the LC50. The approach was described by Kerr and Meador (1996), and is discussed in Appendix M.

Generalized linear models might not be very suitable for routine toxicity tests, which often produce data with one or no partial effects. The models can utilize zero and complete effects, but would appear to rely strongly on the partial effects. Nonlinear models are further discussed in Sections 6.5.2 to 6.5.13.

Additional methods of less immediate interest are discussed in Section 5, including *Time to 50% effect* and use of a *mortality rate model*, the latter technique probably of more interest for research. Other discussion of potential methods is in Appendix M.

Effective Times, Toxicity Curves, and Survival Analysis

All topics in this section are related to the time taken for the toxic material to act on the organisms. At present, these time-related approaches are not in primary use for the testing programs of Environment Canada; however, they have some advantages as extra features in analysis or for possible adoption in the future.

5.1 Median Effective Times

Key Guidance

- *An alternative approach estimates the time required to affect 50% of the organisms in each of a series of fixed concentrations. The median effective times (ET50s), and modelling of those ET50s, can provide higher levels of information and insight, or usefulness in special situations such as short exposures.*
- *Tests designed to estimate the ET50 also provide the data to estimate the EC50, given appropriate choice of concentrations.*
- *There is no simple and convenient software package for estimating ET50s and confidence limits, but it would be useful to develop one.*

Median lethal times (LT50s) have not been used much in recent decades, but in the past they were the standard way of investigating environmental toxicity. Bliss (1937) used logarithmic time series to demonstrate that log-probit transformations were useful in lethal tests. The LT50 was the endpoint in studies of the effects of pesticides on insects (Finney, 1971) and in the classical Canadian work on tolerance of fish and aquatic invertebrates to lethal temperatures, oxygen, salinity, and toxicants (e.g., Fry, 1947; Shepard, 1955; McLeese, 1956). A time-based approach (e.g., LT50) could be helpful in evaluating rapid effects of a dangerous toxicant. For example, it could predict potential damage to fish swimming through a plume of effluent.

To determine acute ET50s, a toxicity test uses a group of organisms at each of several concentrations in a standard logarithmic series. The number of organisms affected in each concentration is observed at successive times which increase in an approximate logarithmic series. For fish, the observation times in hours might be 0.5, 1, 2, 4, 8, 14 ± 2, 24, 48, 96, and perhaps 7 days. For shorter-lived organisms, the time-scale would be adjusted downwards as appropriate.

For a given concentration, the cumulative percent effect is plotted on a probit scale against the logarithm of exposure time. A line would be fitted by eye, and the LT50 read from the graph. When completed for all the concentrations, the findings could be similar to Figure 11, which shows a classical example of mortality times for fish in reduced oxygen (Shepard, 1955). If a suitable range was chosen, severe concentrations would generate short ET50s, and some mild concentrations might only elicit mortalities less than 50% (right side of Figure 11).

The technique could be used for sublethal toxicity, but the effect would have to be easily observed and immediately evident, not delayed. The effect would have to be quantal, or else defined relative to a control, in the same manner as an inhibition concentration (ICp). The term *median effective time (ET50)* is suitable for sublethal effects, as well as for lethal effects.

A series of ET50s could be used to produce toxicity curves such as those in Figure 12. At first glance, the curves appear to be the usual toxicity curves (Section 5.2), but they have *concentration* as the x-axis and time (ET50) as the y-axis. The curves for copper and zinc in Figure 12 appear straighter than usual, with very abrupt thresholds of effect. Below those threshold concentrations (left side of graph), more than half of the test organisms survived for long times; apparently acute lethality had ceased and the organisms were able to deal effectively with the metals.

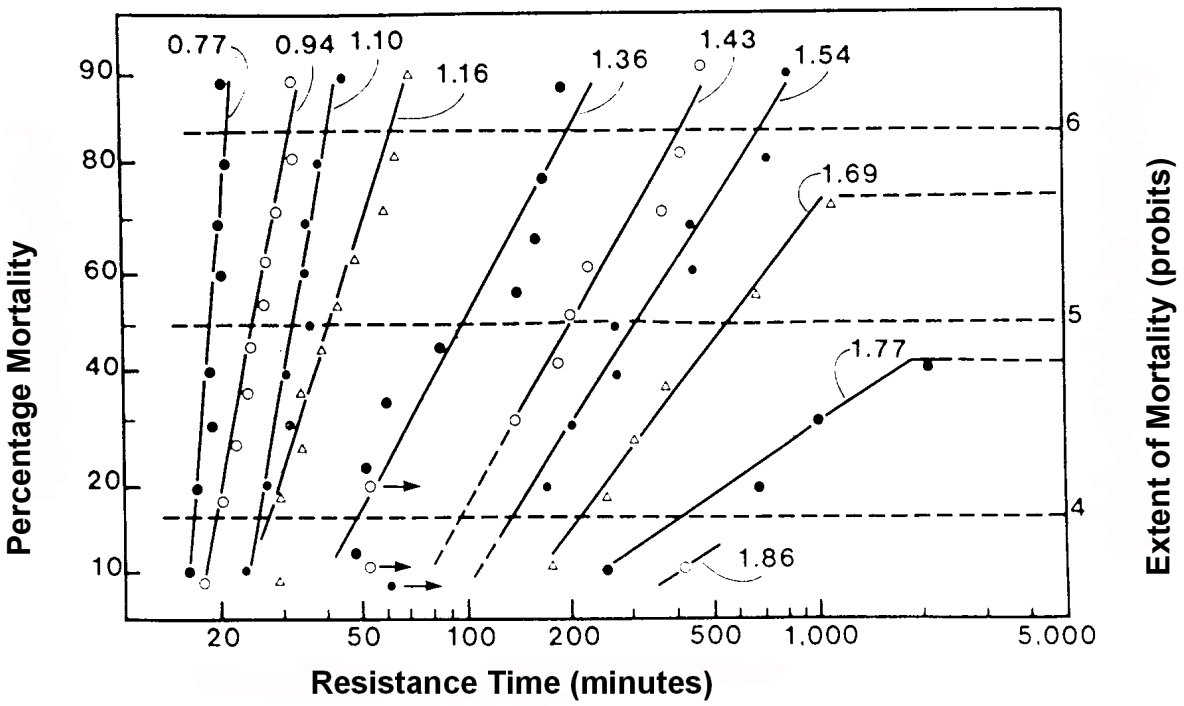


Figure 11 Time-related mortality of brook trout exposed to low concentrations of dissolved oxygen (from Shepard, 1955). Oxygen concentrations are indicated at the tops of the probit lines. The maximum exposure time of 5000 minutes represents about 83 hours. Successive cumulative mortality in each group of fish is plotted on the vertical probability scale, and a straight line is fitted to each. Mortality apparently ceased in the three mildest treatments (right-hand side).

Unfortunately, there is no simple, specially designed computer program for estimating confidence limits for an ET50 [37]. The old method for obtaining these confidence limits is a simplified nomographic procedure (Litchfield, 1949). The standard computer programs for estimating an EC50 and its confidence limits are not valid for estimating an ET50, which arises from repeated observations on the same groups of organisms. Kooijman and Bedaux (1996) offer a program for analysis of toxicity data that might remedy this situation. It is primarily for nonlinear analysis of sublethal quantitative data, but the authors claim that it can also estimate endpoints *with confidence limits*, for effective concentrations (EC50s), and effective times (ET50s). These capabilities have not been verified for the present document because of initial difficulties in operating the program. Mainstream statistical packages (SAS, SPSS, SYSTAT) could estimate the ET50 and confidence limits with relative ease, although they are not exactly "off-the-shelf" methods for a toxicity laboratory.

Using an ET50 as an endpoint for each concentration is predictably more efficient in providing information than using EC50s. In general, about half the information is lost if only the EC50 is estimated. Dixon and Newman (1991) state that "significant statistical benefits accrue from the small amount of additional work" in obtaining time-to-death data, compared to determining the LC50. Similarly, Newman and Aplin (1992) express regret at the lack of attention paid to time-to-effect approaches in

[37] A computer program was written and used at B.C. Research, sometime in the 1970s. It was derived from the method of Litchfield (1949) and apparently worked well (D.J. McLeay, 2004, personal communication, McLeay Environmental Ltd., Victoria, British Columbia). Recent efforts to find this program have not been successful.

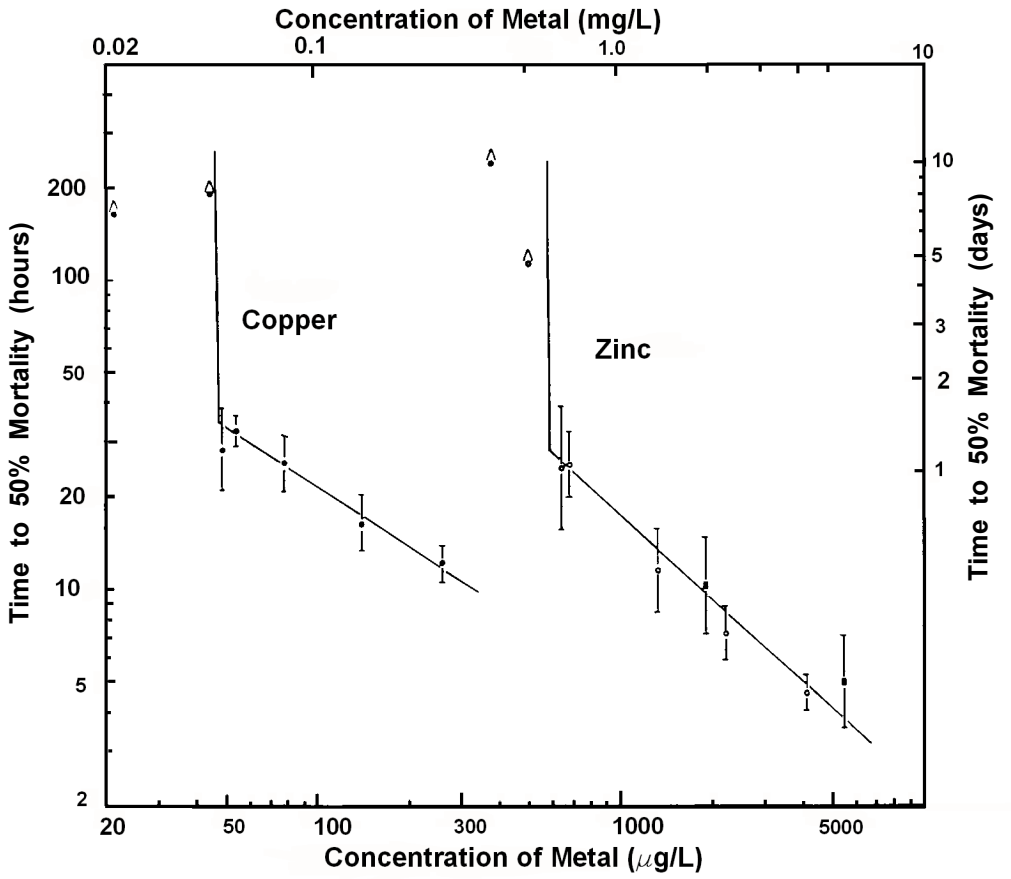


Figure 12 Times of median effect for Atlantic salmon exposed to copper and zinc (from Sprague, 1964). Confidence limits about the ET50s were estimated by the method of Litchfield (1949). Points with vertical arrows represent survival of more than 50% of the test fish during the exposure period indicated by the position on the time axis.

environmental toxicology. They point out that this approach does not lose the standard endpoint (EC50), but gains extra information (the series of ET50s), and enhances interpretation of data (the opportunity to look for meaningful irregularities in the effects).

Examples and further explanation of the gain in information are provided in Bliss and Cattell, 1943; Gaddum, 1953; Sprague, 1969; and Suter *et al.*, 1987. An expected result would be narrower confidence limits on an ET50 compared to an EC50. Another advantage is circumventing the complication of inverted estimates of EC50 and its confidence limits (see Section 9.4).

There would be an even greater gain in information with methods that considered the time course of effects (not just the ET50). There might be additional revelations about what was happening in a toxicity test. Sometimes a "pause" might be noted in the progress of effect, possibly indicative of a change in mechanism of toxic action. Differences in slopes of adjacent probit lines could provide clues about actions of the toxic substance. A break and flattening of a probit line could indicate decomposition or disappearance of the active toxic agent(s). A double bend in the line could indicate two modes of action at shorter and longer times, or the presence of two toxic agents.

One pitfall that should be avoided is any attempt to judge the relative toxicities of different materials, on the basis of short-term ET50s (i.e., effective times in very high concentrations). The comparison can be

very misleading (examples in Sprague, 1969). Similarly, comparisons of EC50s based on short exposure-time (again, involving high concentrations) are misleading. Comparisons are much more meaningful when they are based on times and concentrations that are at or near the threshold of effect (Section 5.2).

Considering all the advantages of ET50s, it is regrettable that methods have swung firmly towards estimating only EC50s. A data-base collected for ET50s such as those in Figures 11 and 12, could still be used for definitive estimates of EC50s. For example, a 96-h EC50 could be estimated in the usual way from percent effects in the various concentrations at 96 hours of exposure. Only the raw observations should be used as input for estimating an EC50; it would not be valid to pick smoothed percent effects from fitted lines such as those in Figure 11.

5.2 Toxicity Curves and Thresholds of Effect

The term *toxicity curve* has a particular meaning in environmental toxicology. It is a graph showing a series of median lethal concentrations plotted against their exposure times, both as logarithms. Alternatively, it could be a series of median lethal times plotted against their exposure concentrations, again as logarithms (Figure 12).

Key Guidance

- *A toxicity curve should be plotted as the test proceeds. LC50s can be estimated at key times during the test, and plotted as a toxicity curve (log LC50 against log time).*
- *The toxicity curve demonstrates any unusual relationships, and whether a threshold of acute action was reached by the end of the test (i.e., the curve became asymptotic to the time axis). An incipient LC50 is a relatively meaningful endpoint since it is determined by the physiology of the test organism rather than by an arbitrary time of exposure.*
- *Most toxicants appear to produce an incipient LC50 in the standard 96-h exposure with fish, and in 10- to 14-d tests for sediment or soil toxicity using invertebrates.*
- *In addition to reporting the EC50 for a standard exposure-time (e.g., 96-h EC50 for fish), reporting an incipient EC50 or the absence of one would increase the practical and scientific value of a test.*
- *Modelling of the data used for toxicity curves has proven profitable in research studies (Section 5.3).*

The main purposes of a toxicity curve are to show any unusual relationships, and *whether* an asymptote with the time axis was attained. Periodic recording of effects during an acute test provides material for the toxicity curve and increases the information gained from the test, which is particularly true for acute lethal tests. Such tests for fish will be used as examples [38].

A major goal in constructing a toxicity curve is to find out whether a time-independent threshold of lethality has been reached (i.e., no further deaths), and if so, whether it occurs early in the test, or slowly. Threshold is used in the sense of half of the fish showing the effect and half not showing it, thus the median fish has just passed the threshold of effect (see glossary). The concentration at which this occurs can be called the *incipient LC50* (or *incipient lethal level, incipient EC50, threshold LC50/EC50*). That is a relatively robust yardstick of toxicity since it marks the concentration that the average fish can *just* deal with, by excreting or detoxifying a chemical as fast as it enters the body. In other words, the incipient LC50 is determined by the physiology of the fish, and is therefore a relatively meaningful and firm endpoint describing acute toxicity.

---

[38] The acute tests with fish are typically four days in duration. For acute mortality of invertebrates in sediment or soil, Environment Canada's tests are usually 10 to 14 days long, sometimes with optional inspection of mortality at 7 days (EC, 1992e; 1997a,b; 1998b; 2001a; 2004a). For the soil/sediment tests, it is not generally feasible to establish a toxicity curve because of the difficulty of establishing mortality at intermediate times, and the possibility of damaging the animals during inspection.

The advantages of comparing results for different exposure times in acute toxicity tests are described by Sprague (1969), Newman and Aplin (1992), and Lloyd (1992). If no threshold is found, it is a warning that effects might continue with extended exposure at very low concentrations.

Estimates of EC50s can be made as the experiment proceeds (e.g., 4, 8, 24, 48, and 96 hours of exposure), and these can be plotted as a toxicity curve using logarithmic scales (Figure 13) [39]. It may become evident that the curve becomes asymptotic to the time axis, i.e., acute lethal action has ceased (right-hand side, Toxicant A in Figure 13). It is of considerable interest to know whether there is a low concentration which the average organism can deal with during an acute exposure; the remaining organisms would apparently survive the exposure. There is no particular rule for determining whether such an incipient LC50 has been achieved, so the toxicity curve should be interpreted subjectively [40]. Sometimes there can be a very sharp threshold, leaving little doubt about the interpretation (Figure 12).

Even if a short exposure time did not cause 50% mortality, it can still contribute to shaping the toxicity curve. For that exposure time, the LC50 would be higher than the highest concentration tested; a point can be plotted with an arrow pointing up to higher concentrations than those tested (left side of curves in Figures 13 and 14). The fitted curve can miss some points (smoothing) because each LC50 has potential variation (confidence limits).

It would have been desirable to prolong the test for Toxicant B in Figure 13 to see if a threshold (asymptote) was eventually achieved. Therefore, a rough plot of the curve should be made as a test proceeds, for guidance on terminating the test. Even lower concentrations would apparently have killed the organisms with longer exposure. It would obviously be of interest to know of such a situation, which would represent a dangerous type of toxicant, because lower and lower concentrations might cause an effect, given a sufficiently long exposure time.

The use of logarithms of time and concentration in plotting the toxicity curve is of utmost importance, for reasons discussed in Sections 2.3 and Appendix D. A toxicity curve plotted with an arithmetic scale of time distorts the curve and can be highly misleading. One primary error could be that a threshold appeared to be reached at long exposure times, when in fact there was no threshold. *With a graph using arithmetic time, any test could be made to show an apparent threshold, even if one did not exist, simply by running the test long enough.*

A hypothetical example of incorrect axes is shown in Figure 14, in which the upper panel has arithmetic scales of both concentration and time. The curve appears to reach a reassuring asymptote after exposure of 7–10 days (168–240 hours). However, a proper logarithmic plot of the same data, in the lower panel of Figure 14, shows regular continuing mortality and a straight-line relationship with no threshold. In other words, by using arithmetic axes, the investigator would be misled into believing that a toxicant possessed a threshold, below which the toxic effect disappeared, when in fact, the lower concentrations caused toxicity under exactly the same time-concentration relationship as at higher concentrations.

Contributing to the misinterpretation of data used in Figure 14, would be the failure to maintain a regular increase of exposure. The important change in exposures is the *ratio* between successive exposure levels, not the absolute increase (Section 2.3). In most of the test shown in Figure 14, successive exposure times represented a doubling, or close to that. The final pair of inspections represent a three-day interval (from 7 days to 10 days) which might seem like a relatively long interval, and indeed, plots as a long interval on the arithmetic scale. However, it represents an increase of only 1.4 times, and therefore allowed less opportunity for change in observed effect, than did the earlier doubling changes, such as from 1 to 2 days, and 2 to 4 days. This type of

---

[39] An arithmetic scale of concentration would be used instead of a logarithmic one, if the "toxic agent" being studied was temperature or pH, which is already logarithmic.

[40] No standard method of statistical testing has been established to determine whether an asymptote had been obtained. It seems unlikely that any simple method will be provided, partly because of the non-independent repeated observations on the same groups of organisms.

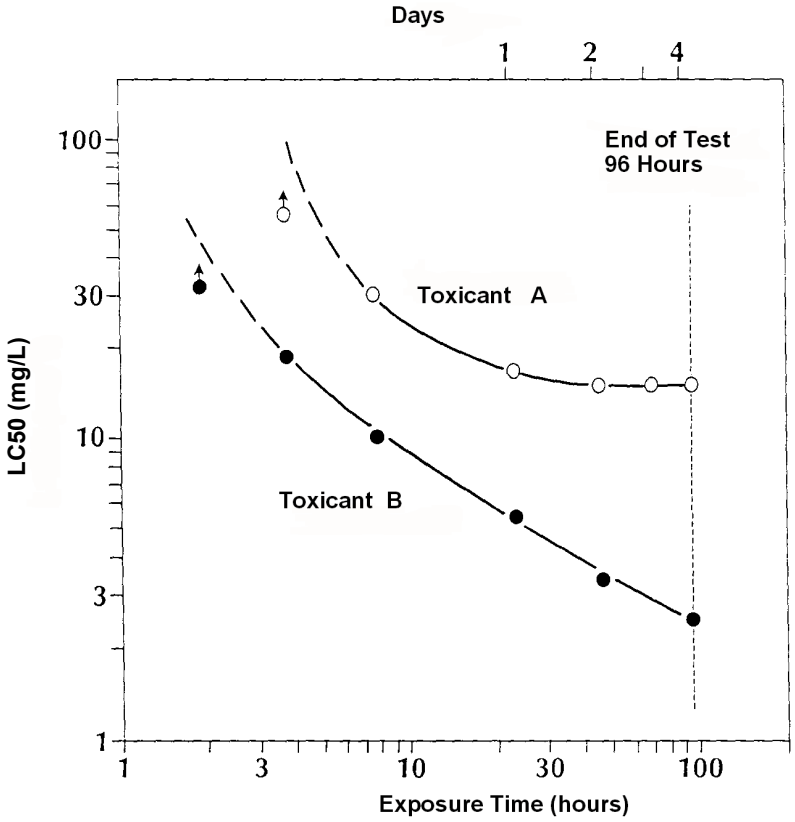


Figure 13 **Toxicity curves for two hypothetical toxicants.** Curves were obtained by fitting lines by eye to all of the LC50s. Logarithmic scales are used for time and concentration. Toxicant A reached an incipient LC50, because the curve became asymptotic to the time axis after about two days. Toxicant B did not reach an asymptote.

mistake seems to be establishing itself in acute tests of soil toxicity (Lanno *et al.*, 1997) and should be remedied.

Estimating the incipient EC50. It is not appropriate to report an incipient EC50 which has been estimated by eye from a toxicity curve. Instead, the curve is used to determine an *exposure time that appears to be in the asymptotic region*, and a final (incipient) LC50 is formally calculated for that time, using a standard technique (Section 4.5), that provides an accurate EC50 with 95% confidence limits.

The OECD (2004) advises against using toxicity curves: "This is not a proper method and should be avoided". However their arguments are not well taken and their main statistical objection is that the "dose-response data at different time points are not independent ...". That would not appear to be a problem since the toxicity curve is merely an informal way of visualizing when acute effects seem to have ceased. As recommended here, the definitive calculation of the incipient EC50 is done in a standard way, completely independent of any data on effects at earlier times. A toxicity curve can add a great deal of insight for understanding toxic effects in a test, and investigators should not be dissuaded from using this tool, by the comments of the OECD.

Designing a test with a toxicity curve in mind might require that additional low concentrations should be tested. The benefit, however, is that a plotted toxicity curve will usually increase understanding of the toxicant's hazard. In aquatic testing with fish, most toxicants produce an incipient LC50 in the standard 96-h exposure (Sprague 1969), while a threshold seems likely in 14-d soil tests with earthworms (Lanno *et al.*, 1997).

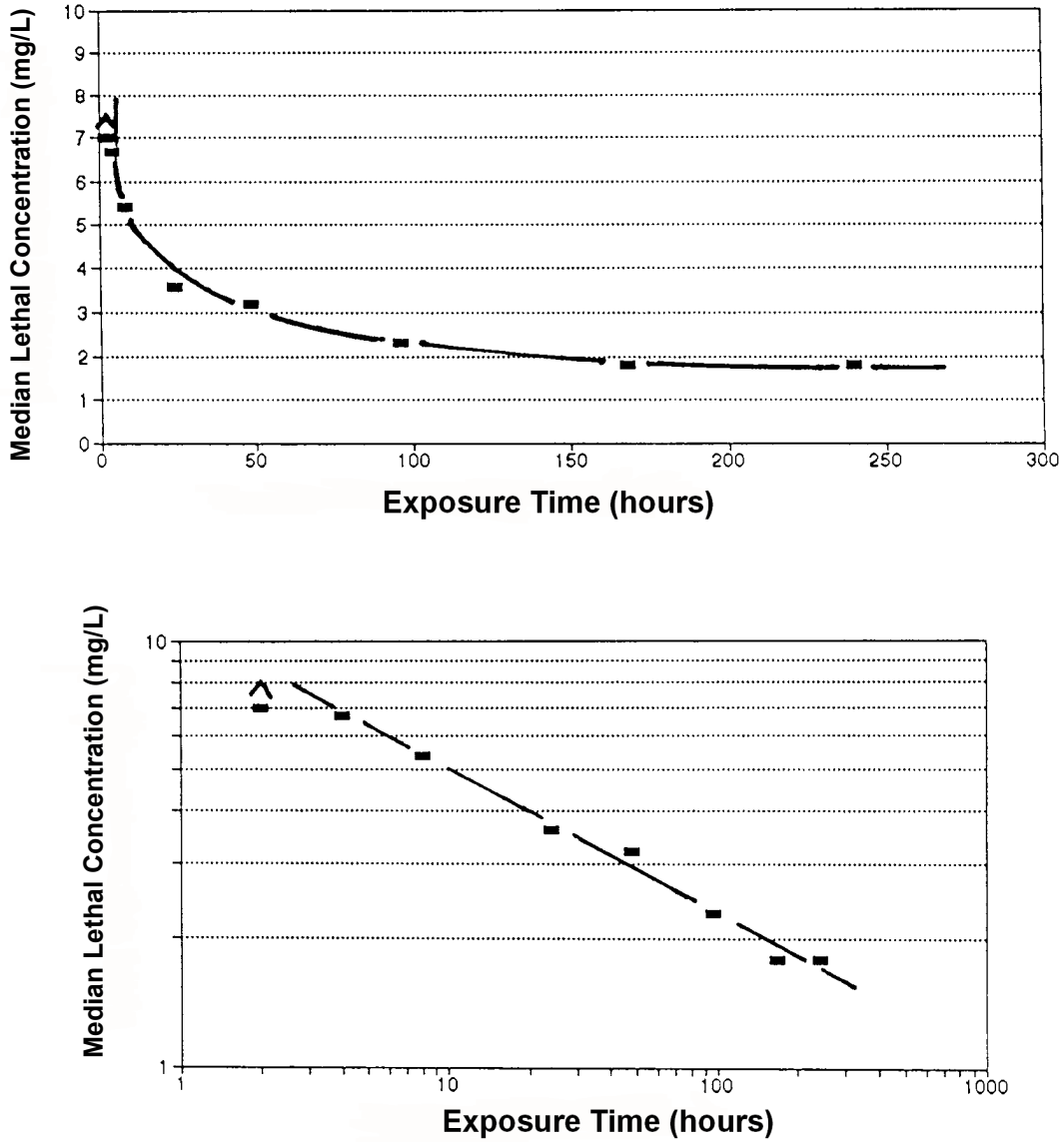


Figure 14 Improper toxicity curve on arithmetic axes. For the arithmetic axes in the upper panel, the curve flattens out parallel to the time axis on the right-hand side. An investigator would be misled to believe that a threshold of acute toxicity had been achieved, so that toxicity would not occur at lower concentrations. In the lower panel, the data are replotted with correct logarithmic axes, resulting in a straight-line toxicity curve. No threshold is evident, and acute toxicity seems likely to continue to lower concentrations, a dangerous type of toxicant. Hypothetical data.

For Environment Canada tests, an EC50 should be estimated for the standard exposure time stipulated in the methods document, such as 96 hours for fish, or 14 days in soil tests with earthworms. If that standard EC50 also represented an incipient EC50 as described previously, that should be reported. If an asymptote was obtained only after longer exposure, a second, incipient EC50 should be estimated for that longer time and reported as an additional meaningful endpoint. It is beneficial to submit a toxicity curve as part of any report on a lethal test. If there is no asymptote observed, an investigator should point that out; *apparent absence of a threshold is of considerable toxicological importance.*

There appears to be some renewed interest in time-related modelling of toxic effects, as indicated in following sections.

5.3 Modelling Effect-times and Toxicity Curves

Statistical modelling of results is not part of Environment Canada's standard tests, so the topic is not extensively examined here. Some mention is made for investigators who might wish to proceed further in analyses of their tests.

There are a few publications on statistical descriptions of toxicity curves. A pioneering study by Alderdice and Brett (1957) fitted a rectangular hyperbola to lethality data for Canadian pulp mill effluent. An incipient LC50 was derived. Carter and Hubert (1984) developed a generalized polynomial equation (growth-curve type), using a multivariate linear model. This was incorporated into a BASIC computer program by Hong *et al.* (1988). They used the program to describe a 14-day toxicity test with fish, producing a three-dimensional graph, time-independent LCps, and toxicity curves with confidence bands. The program has not received wide use. It had deficiencies of not allowing for control effects, and of modelling with arithmetic values of time, so the curves gave a distorted impression of toxicity relationships.

Heming *et al.* (1989) used time-related analyses in an excellent study of the effects of the insecticide methoxychlor, on several species of fish. They were successful in demonstrating several fits for standard toxicity curves. Four models gave good descriptions of the curves, out of eight models tried. Kooijman and Bedaux (1996) offer a complete program *(DEBtox)* for analysis of toxicity data. The program has options for analysis of data on EC50s and median effective times (ET50s), with confidence limits and consideration of the time of response. Others have used a plot of a fitted survival model, to show a three-dimensional relationship between concentration, time, and percent effect (Newman and Aplin, 1992).

Periodic attempts were made some decades ago to extrapolate toxicity curves for lethality, to predict threshold toxic effects, including sublethal effects. Lee *et al.* (1995) revived this quest in a sophisticated way by developing three models to predict chronic lethal effects in fish. They applied multiple regressions to data on acute lethality, with choices of data transformed to log concentration and log time, reciprocal of time, or log of reciprocal of time. Trials against 28 sets of data showed that predicted values were generally close to observed chronic lethality, and at least of the same order of magnitude. The practical application of the method would be using inexpensive acute tests to identify dangerous pollutants that deserved study by more expensive chronic tests.

5.4 Analyses of Survival over Time

Key Guidance

- *Mortality/survival rates and analysis represent a group of advanced statistical procedures for examining toxic effects. They are well-known in biomedical research, and recent literature shows relevance for environmental toxicity. The research techniques would need to be tailored for routine use by investigators.*
- *The "repeated measures" procedures of statistics might often be appropriate for analysis of repetitive test observations.*

5.4.1 Mortality Rate

Mortality and survival represent two sides of a coin, but Borgmann (1994) developed an approach in environmental toxicology, which integrates the effects of time and concentration under the name *mortality rate*. The approach could be beneficial in research, particularly for long exposures which combine observations on mortality with sublethal effects such as weight. It would be useful for long-term tests with sediment and invertebrates, in which there is often continued mortality. It is also advantageous when there are few concentrations with partial effects. Investigators interested in pursuing this approach could gain an appreciation of its methods and applications in Borgmann (1994).

Although the mortality rate is a continuous or quantitative variable, Borgmann (1994) uses it to

integrate mortality, which is a quantal effect. The mortality rate model starts from the different assumption that all of the test organisms have the same sensitivity to the toxic material, and that mortality is a random event which can be quantified as a rate. Total mortality rate can be statistically separated into components of control mortality and toxicant-induced mortality rate. A concentration-effect curve can be produced, and the LC50 estimated. The approach can also be used for production rate of biomass. Guidance on working with rates is given in a textbook by Fleiss (1981).

5.4.2 Survival Analysis

A particular group of techniques is signified by the name *survival analysis*, often used in biomedical studies. These are well-established and profitable methods of examining time-related toxic effects, although they are somewhat complex (Newman and Aplin, 1992). A brief but excellent introduction is given by Crane and Godolphin (2000). They provide examples and access to the literature for such topics as *two-step linear regression, multifactor probit regression, survival-time modelling*, and *kinetic models*. The kinetic approach includes more-or-less theoretical consideration of the behaviour of toxicants in living organisms, with the possibility of better determining incipient toxic concentrations and *true* no-effect concentrations (NEC).

Heming *et al.* (1989) applied these techniques in their applied consideration of pesticide toxicity (see Section 5.3). Another good example of survival-time modelling is provided by Newman and Aplin (1992), who analyzed salt toxicity to a freshwater fish. They did standard analyses for LC50s, but showed that survival-time modelling provided much more information. Their methods allowed median survival times at any given concentration of toxicant, low levels of mortality such as 5%, and toxicity for a given body weight of fish to be predicted, all with accompanying estimates of standard errors. Newman and Aplin (1992) recommended the SAS procedure LIFEREG for these analyses.

Among the strongest proponents of these more-sophisticated analyses for environmental toxicology are Kooijman and Bedaux (1996; also Kooijman, 1996). A comprehensive introduction to these advanced topics, for those with some statistical skill, is provided in a recent book by Crane *et al.* (2002). Chapter 5 of the book shows the advantages of survival-time modelling compared to conventional probit/logit analyses of acute lethality. The book proceeds to more advanced techniques of time-related analyses, such as life tables and exponential survival functions. Dixon and Newman (1991) point out that analyses of effect-times are "readily implemented with several common software packages" including SAS and SYSTAT, but that does not represent an easily accessible program tailored to the needs of all toxicology laboratories. Another source of information on survival analysis is Parmar and Machin (1995).

5.4.3 Repeated Measures

Repeated measures is the name applied to procedures and analyses based on measurements collected over time from the same source. If a blood sample were taken from a fish on several occasions, that would yield repeated measurements on a *sampling unit*. If measurements were made on aliquots of an algal suspension extracted from a larger vessel over time, the repeated measurements would be made on the *experimental unit.* (These would not be *subsamples*, which would be collected simultaneously.) The approach is not often used in environmental toxicology and changes in effect over time "can and often should be analyzed using repeated measures and related approaches, but those approaches may be more complex" than the design which is shown in a table for ANOVA (Paine, 1996). There is a need for someone to establish a pattern for using these more sophisticated approaches to time-effect data in environmental toxicity studies.

Point Estimates for Quantitative Sublethal Tests

Estimating the endpoints of sublethal tests is a major interest in environmental toxicology. Four of the nine topics discussed by Canadian environmental toxicologists at the Quebec City meeting, were specifically focused on determining sublethal endpoints (Miller *et al.*, 1993).

This section starts with orientation on choosing endpoints and general items for all sublethal tests and proceeds with specific coverage of *quantitative point estimates* that can be used to describe sublethal effect. Sublethal tests get further coverage in Sections 7 and 8.

6.1 General Items on Sublethal Tests

In quantitative toxicity tests, the investigator does not merely observe whether an organism shows an effect or does not show it, but instead makes quantitative (continuous) measurements. The weight of each test organism might be measured in grams, the number of progeny might be counted, the activity of an enzyme might be measured, etc. Whole-organism effects are of great practical interest, and are considered in this document. The effects commonly measured are attained size, degree of larval development, fertilization, germination, and number of young produced. In a few cases, the sublethal effects are quantal but can be treated as quantitative because of the large numbers of observations (see following text).

The quantal methods described in Sections 4 and 5 are neither appropriate nor valid for quantitative measurements, and no attempt should be made to apply them. However, mortality might sometimes be an additional measurement in a test designed to show sublethal effects, and quantal analysis would be appropriate for that mortality data in dual-effect tests (Section 8).

There is a choice of approaches and methods in sublethal tests, and some comments on that choice are made here. Some general items are also covered here since they apply to both point estimates (this Section) and hypothesis testing (Section 7).

6.1.1 Types of Tests and Endpoints

Key Guidance

- *Environment Canada has published a variety of sublethal methods for testing water, sediment, and soil, using chronic, sub-acute, or acute exposures.*
- *Most of the tests involve quantitative effects, e.g., measurement of the weight of organisms. Some quantal effects could also be measured in the same test, such as mortality after long exposure, or mortality of the first generation of earthworms.*
- *A point estimate is recommended as the best quantitative endpoint. This is usually a specified degree of reduction in performance compared to the control, most commonly 25% impairment in Environment Canada tests. An example would be the concentration associated with a weight that was 25% lower than the control.*
- *Unsatisfactory methods of analysis have been commonly used for making point estimates. The method of interpolation is easy but neglects much of the data. More sophisticated methods using linear and nonlinear regression are becoming more widely used, and are now standard in Environment Canada's new methods for soil toxicity. The methods require that laboratory staff understand the judgements in selecting appropriate mathematical models.*
- *Hypothesis testing is commonly used to identify concentrations with significant effects compared to the control. This method has many flaws and is not recommended for future use (see Section 7).*

Types of Tests. In recent years, Environment Canada has produced standard methods for a number of sublethal toxicity tests, mostly for free-living aquatic organisms and sediment-dwellers. Further methods for sediment and soil toxicity are under development. The tests are catalogued in Appendix A, and are briefly listed here to indicate the wide range of organisms and sublethal effects (EC, 1992a–f; 1997a,b; 1998a,b; 1999b; 2001a,b; 2002a; 2004a–c). Some of the toxicity tests use hypothesis testing for analysis, but quantitative point estimates are recommended for most of them. Some tests are dual-effect (Section 8), and are so marked in the list. The second effect is often quantal, usually mortality which is obviously not sublethal.

Organisms	Type of Test
Marine luminescent bacterium, *Vibrio fischeri*.	Sublethal functioning in liquid as indicated by amount of light produced.
	Sublethal functioning in sediment.
Freshwater green alga, *Pseudokirchneriella subcapitata [*formerly *Selenastrum capricornutum]*	Inhibition of growth and reproduction as indicated by number of cells.
Freshwater aquatic plant, *Lemna minor*.	Inhibition of growth.
Terrestrial plants.	Emergence and growth of plants exposed to contaminants in soil.
Marine and estuarine polychaete worms.	Inhibition of growth in sediment and mortality (dual effect).
Earthworms in soil.	Avoidance behaviour. Number and growth of progeny, mortality of first generation (dual-effect).
Springtails.	Number of progeny and mortality of first generation (dual effect).
Marine echinoids, sea urchins, and sand dollars.	Fertilization success after initial exposure of sperm, continued after adding eggs.
Freshwater crustacean, the waterflea *Ceriodaphnia dubia*.	Number of young produced, and long-term mortality of the adults (dual-effect).
Marine/estuarine crustacean amphipods	Apparent avoidance of sediment, ability to burrow and rebury, and 10-day mortality (dual-effect).
Freshwater crustacean amphipod *Hyalella azteca*.	Growth (weight increase) and mortality in sediment after a 14-day exposure (dual-effect).
Freshwater midge larvae (Chironomidae) *Chironomus tentans* or *C. riparius*.	Growth (weight increase) and mortality in sediment after a 14-day exposure (dual-effect).
Cyprinid freshwater fish, the fathead minnow.	Growth of newly hatched larvae, and their mortality (dual-effect).
Freshwater salmonid fish.	Success in development of embryos, embryo/alevins, or embryo/alevins/fry.

Quantal endpoints. A test designed to measure sublethal effects might also have mortality as one of several effects. There might be short-term mortality at higher concentrations. Chronic exposure might have various sublethal effects that would finally cumulate and cause death. Analysis of mortality should be done by probit regression or other quantal method (Section 4).

There are a few quantal sublethal tests. One measures avoidance of contaminated soil by earthworms (EC, 2004a). The analysis would produce an ECp, using the same quantal procedures as for LC50 (Section 4). Two other tests measure fertilization success with gametes of rainbow trout (EC, 1998a) and echinoids (EC, 1992f). The effect is quantal, but an alternative analysis can be used for the echinoids, as described in the following text.

Quantitative estimates on quantal data. If the number of quantal observations (organisms) is large, $\geq$100 in a replicate, it is acceptable to analyze the data as if they were quantitative. An example would be the echinoid fertilization test (EC, 1992f), which has counts of 100–200 eggs per container. The eggs are classified as fertilized or not fertilized, i.e., quantal

data as indicated previously. Because of the large numbers, however, the change in percent effect caused by one individual reacting, would be low enough that these data can be treated as if they represented a continuous distribution [41]. Environment Canada recommends estimating the ICp, a quantitative endpoint, in the echinoid test. There is an optional extra of hypothesis testing, although it retains all the disadvantages listed in Section 7.1.2. Another example is in tests of algal growth/reproduction, in which the basic variable is the number of cells, which is quantal. Because there can be thousands or tens of thousands of cells, the distribution of numbers can be considered continuous and the test is treated as a quantitative one.

On the other hand, Environment Canada's early life-stage test with salmonids has only 40 eggs per container for a total of 120 per treatment (EC, 1998a). The test yields quantal data (viable/non-viable eggs), and the required endpoints are EC25 and EC50, appropriate for these numbers of individuals. The numbers in the containers are not high enough to treat the data as if they were quantitative.

Quantitative point-estimates. The preferred quantitative endpoint in sublethal tests is called a *point-estimate*, which is a specific point on the continuous scale of concentration (see Section 6.2.2 for a list of advantages). Usually the endpoint is chosen to represent a certain degree of reduced performance compared to the control, e.g., 25% fewer progeny than the number produced in the control. The method has a basic assumption that there is a reasonably regular dose-effect relationship that can be used to estimate the single endpoint.

There are two main problems with using a point-estimate.

- First, selection of an appropriate degree of impairment is clearly a subjective choice by the investigator or by consensus of the profession (should it be 25% impairment, or 10% as is fairly common in Europe?). Choosing the greater degree of effect (25%) will make it more clearly the result of the test material and not merely experimental variation. A lower level of effect (e.g., 10%) will mean that the endpoint is closer to a truly "safe" concentration (see Glossary and Section 6.2.4).

- Second, the distributions of effect assume a variety of shapes, and accordingly, require a variety of mathematical models to describe them. However, real progress has been made in developing a standard approach, which starts with selecting a suitable model from a small suite of choices (Section 6.5.8).

An investigator considering suitable methods of analysis could follow Figure 15 downward to the left through "Point estimate", and find two general choices of method, described in Sections 6.4 and 6.5.

- First is the nonparametric method of ***Smoothing and Interpolation***. This has been the usual method for analysis, but it has several deficiencies and deserves to be replaced (Section 6.4.1).

- A second choice is ***regression***, either ***linear or nonlinear***, suitable for a variety of dose-effect distributions. General-purpose statistical programs such as SYSTAT can be used in a standard analytical approach, now adopted by Environment Canada (Section 6.5.8). The toxicological program CETIS also offers mathematical models for nonlinear regression. There is a continuing requirement for some mathematical knowledge, to select the appropriate nonlinear model and carry out the mathematical treatment.

---

[41] The quantal (binary) data will predictably adhere to a binomial distribution, and appropriate statistical analyses would use methods for that distribution such as the chi-square test. However, with the large numbers, the distribution of observations comes to resemble a normal distribution. Little bias or error is introduced by using quantitative statistical techniques to obtain an endpoint.

For example, if 10 eggs were in a replicate, and eight were found to be fertilized, then each egg had an influence of 10% on the results, from 70% total effect if that egg had been unfertilized, to 80% total effect if it were fertilized. That 10% jump is an abrupt and appreciable change, indicative of the quantal nature of the data. However, if there were 100 eggs in the replicate, each egg could influence the overall result by only 1%, say from 70% to 71%. For all practical purposes, that represents a quantitative effect.

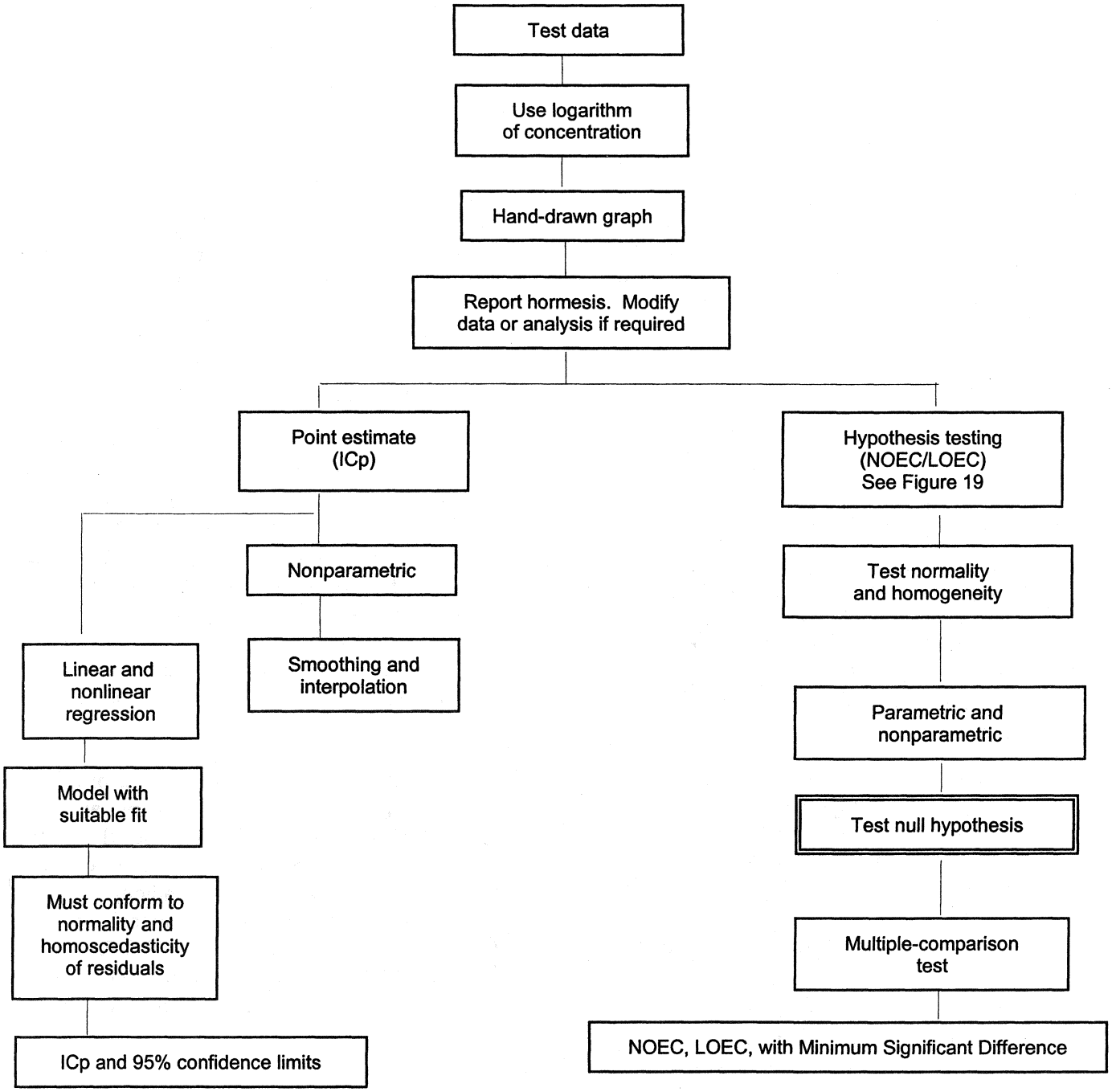


Figure 15 Analytical sequence for multi-concentration quantitative toxicity tests.

Participation of a statistician in design and analysis (Section 2.1) is especially important in sublethal point-estimates. Some of the more sophisticated methods towards the end of Section 6 have an absolute requirement for qualified statistical advice.

Hypothesis testing. This alternative to point-estimates has been commonly used, and is allowed but no longer recommended in various new test methods of Environment Canada. The approach determines the lowest concentration which caused a statistically significant effect in the test (the LOEC; right side of Figure 15). It is described with its deficiencies in Section 7.

6.2 Elements of Sublethal Point Estimates

Key Guidance

- *The ICp (Inhibiting Concentration for a specified percent reduction of performance) is the standard endpoint for quantitative sublethal tests. The value of "p" in ICp is commonly 25% or 20%, or sometimes 10% in Europe. It is selected by biological judgement, and has no statistical root.*
- *Europeans and some groups in the USA often call this endpoint the ECp, a misleading error since that refers to quantal tests in which a specified proportion of organisms show a particular effect.*
- *The ICp brings many advantages. It is a single concentration, confidence limits can be calculated, and variability of data should not systematically influence its value. Disadvantages are fewer and minor.*
- *Replication might not be required by a test document, but even modest replication is beneficial. It can help distinguish between (a) variability within the test, and (b) deviation about the model selected as the dose-effect pattern. Extensive replication is needed when determining endpoints with nonlinear regression.*
- *Before formal mathematical analysis, a hand plot of results should be made to inspect the general form of the dose-effect curve, and to provide a rough endpoint to check the computerized estimate.*

6.2.1 Terminology

The name for a quantitative sublethal point estimate in North America is *ICp*, signifying *the Inhibiting Concentration for a (specified) percent effect*. It is the concentration that is estimated to cause a designated percent impairment in a biological function, compared to the control. For example, an IC25 could be the concentration estimated to result in 25% fewer progeny per brood, relative to the control.

Quantitative sublethal effects should not be described with the terms EC25, EC50, etc.; those terms are for quantal data *(effective concentration for a specified percentage of individuals)*. Referring to 25% of individuals being affected (EC25) is quite a different thing than referring to a performance that is 25% "worse" than that of the control (IC25). Using the correct term provides information to others about the type of test, type of data obtained, and appropriate type of analysis. Use of the incorrect term is misleading.

Particularly disturbing is the incorrect use of EC50 in Europe, even by technical people of some reputation, and especially by working groups of the OECD and ISO. This mistake also occurs in North America, in some statistical packages (CETIS), among mathematicians (surprisingly), and notably for tests of bacterial luminescence. Even the USEPA sometimes fails to draw a clear distinction between quantal and quantitative tests when describing point estimates (USEPA, 1995).

Other names and symbols have been suggested for endpoints of particular statistical techniques, but ICp seems suitable for all quantitative estimates.

6.2.2 Advantages of Point Estimates

The chief advantages of point estimates are that a single concentration is obtained as the endpoint and confidence limits can be estimated. A list of other advantages follows, in comparison to the alternative approach of NOEC/LOEC. Most of the items assume that the point estimate was obtained from a regression. Similar lists have been provided by

Stephan and Rogers (1985); Pack (1993); Noppert *et al.* (1994) [42]; Chapman, (1996); Moore, (1996); OECD (1998), and others.

(a) A single concentration is designated as the endpoint.

(b) The endpoint can be any concentration within the range tested, and does not have to be a concentration selected by the investigator and used in the test.

(c) Confidence limits can be placed on the endpoint. Other customary expressions of variation can be calculated, such as the standard deviation.

(d) The value of the endpoint would not usually be biased in a consistent direction by the degree of natural variation, by variation induced by the investigator's care in conducting the test, or by the number of replicates (the precision, however, could be influenced).

(e) Choice of α, the level of statistical significance, does not affect the estimate of toxicity.

(f) If the endpoint is estimated by regression, the particular method chosen would usually have a relatively small effect on the estimated endpoint, at least for central values of "p" in ICp.

(g) Use of ICp encourages the consideration of degrees of impairment in the real world, and discourages thoughts that the NOEC derived by hypothesis testing is a biological "no-effect" level.

Disadvantages of point estimates can also be listed. Some of these simply indicate problems to be solved, or methods which still have to be standardized.

(a) The magnitude of effect for the endpoint (the "p" in ICp) is not an absolute, and requires subjective input and concordance among investigators.

(b) Precision of the estimate of the endpoint depends upon the number of test concentrations, their numerical values, the number of replicates, and on the selection of an appropriate mathematical model for the relationship. Thus, the choice of concentrations may influence the estimated ICp, particularly for low values of "p".

(c) Confidence limits become wider for smaller values of "p" in ICp.

(d) The model chosen to fit the data can influence the value estimated for the endpoint, particularly again, for endpoints of low effect.

6.2.3 Replicates

Section 2.5 provides full information on replication for point estimates. For regression, only one measurement at each concentration is an absolute requirement for estimating the endpoint and confidence limits. However, replicates are needed if the investigator wishes to choose among linear and nonlinear models to fit the data and assess the goodness of fit. For less desirable estimation of an endpoint by smoothing and interpolation (ICPIN, Section 6.4), there must be at least two replicates for each concentration to calculate confidence limits, and five or more are desirable. Environment Canada documents usually recommend three replicates for point estimates, in case they are needed for hypothesis testing, but four would be needed for certain methods of nonparametric analysis.

6.2.4 Selecting the Degree of Effect for an Endpoint

The choice of a value for "p" in ICp is entirely arbitrary. It is a judgemental decision made by investigators, not something that is governed by mathematics. In North America, there have been some informal attempts to establish IC20 as a standard endpoint in aquatic tests, but IC25 is more frequently used than other values (i.e., 25% reduction in performance).

There has been some justification of the IC25 as being similar, in many cases, to the NOEC [43]. That is

[42] Alternatively, see deBruijn and Hof (1997), van der Hoeven (1997), and van der Hoeven *et al.* (1997).

[43] The evidence is not copious. One major comparison for aquatic tests indicated that the IC25 was similar to the NOEC for 23 effluents and reference toxicants, in sublethal tests with sea urchins, sheepshead minnow, and the red macroalga *Champia* (USEPA, 1991a). A set of tests with daphnids and a single reference toxicant also showed similarity of the two measurements (OECD,

not a particularly good argument in view of the many deficiencies of the NOEC/LOEC approach (Section 7). The relationship between ICp and NOEC could change with the power and variation of a particular test, with the effect measured, and with the material tested.

Europeans have found that estimation of IC10 is feasible in several types of tests. It has been used to describe inhibition of algal growth (ISO, 1999) and is sanctioned for other methods of the International Organization for Standardization (ISO, 1998). Certainly the IC10 is an acceptable endpoint in tests, if it can be estimated with suitably narrow confidence limits, and certain other conditions are satisfactory (see following text). Promoting the IC10 as a potential endpoint might help to replace the less desirable NOEC/LOEC. Some clients of testing programs, among industry and environmental organizations, feel that the endpoint of a test should sound "safe", and NOEC has that sound. The IC10 is obviously closer to a no-effect level than is the IC25, and gives a stronger impression of a low and relatively protective endpoint.

The Statistical Advisory Group of Environment Canada listed various influences in choosing a suitable value of "p" (Miller *et al.*, 1993), and these are given here in abbreviated form.

- The basic question is whether to choose the value of "p" in ICp on the basis of ecological significance, or for statistical convenience.
- A low value of "p" is desirable from the biological point of view, to obtain a sensitive estimate of toxic action.
- A low value such as IC10 would mean working at one end of a dose-effect relationship, perhaps resulting in undesirable variation of the estimate. An IC50 would be statistically desirable, but biologically, a lower value of "p" would be required in sublethal tests.
- The option of a low value of "p" will be specific to the type of test and the effect measured. If a variable effect were being measured, an IC10 might well be within the zone of normal biological variability, resulting in uncertain interpretation of the endpoint. The variability observed in the control should play a part in selecting the value of "p".
- IC25, or sometimes IC20, seems to have gained favour in North America and some other countries, as a good minimum for indicating a "biologically significant" change.

The "p" of ICp should have a value that is greater than whatever value is specified in the methods document as the upper limit of acceptable effect in the control and this qualification should be added to the second last point. The IC10 would generally seem to be a practical lower limit as a dependable endpoint. Considerations on choosing a "p" for ICp are somewhat parallel to those for quantal endpoints (ECp) mentioned in Section 4.2.5.

Statistically, there is a definite situation of wider confidence limits on the ICp as "p" becomes lower. For very low values of "p", it might be difficult to obtain a reasonable estimate of a concentration with acceptably narrow confidence limits. Part of this effect is because the ICp and its limits might be estimated by *inverse regression*, in parallel fashion to quantal estimates of ECp (see Section 9.4 [44]). Working in the tails of the regression, the limits

1997). A compilation by Suter *et al.* (1987) of 176 sublethal tests with fish indicated that the IC25, on average, was almost equal to the TOEC (a higher concentration than the NOEC). However, even the *mean* ratios IC25/TOEC, for various effects on the fish, varied from 0.5 to 3.2. Later, Suter *et al.* (1995) concluded that a sublethal inhibition of 20% to 25% was about the lowest that would correspond to a statistically detectable effect (LOEC).

[44] Briefly, the investigator chooses a set of concentrations and observes the effects at each of them. If a regression is fitted, the effect is the dependent variable, and log concentration is the independent variable. The confidence intervals of the regression are in terms of the *effect*, and as always, they are wider at the extreme concentrations than in the "centre". The investigator wishes to reverse the interpretation, with confidence limits expressed in terms of concentration about an endpoint (concentration) which is estimated to cause a specified effect (p). This can be considered an "inverse regression". The inverted limits are asymmetric, and in the lower part of the regression, the lower limit can become particularly wide (similar to Figure 7). See Section 9.4.

become large, and for concentrations at the low end, it is not uncommon to have the lower limit go to infinity when inverse regression has been used.

The choice of concentrations can reduce this problem. Since confidence intervals always flare out from the mean of the independent variable, a good design for the experiment would mean that the independent variable (concentration) would be centred on the p-value of interest. Concentrations, therefore, should be selected so that they were near the endpoint of interest, say the IC10. Of course that is a little difficult to foresee, and there are conflicting priorities (Section 2.2), but the principle should be kept in mind.

General experience indicates that the IC10 is less desirable in the tests commonly done with the methods of Environment Canada because it can experience wide limits. The selection of IC25, as has become customary for Environment Canada and more broadly in North America, or the IC20 as an alternative, has merit as an achievable endpoint that is still meaningful.

6.2.5 Selecting the Biological Variable for an Endpoint

The *effect* that is to be analyzed might influence the ICp, moving it higher or lower. Accordingly, the choice of effect might be an important influence in selecting the value of "p" in ICp. This could be particularly important in dual-effect tests, a topic considered in Section 8.

6.3 General Steps in Estimating a Sublethal Endpoint

Key Guidance

- *A rough hand-plot of the data should be the first step in analysis. This shows the general nature of the results and provides a check on the final estimate of the endpoint.*
- *Linear or nonlinear regression is the method of choice. If that cannot be done, the alternative is to fall back on a common method using interpolation (the "ICPIN" procedure).*

6.3.1 Plotting Data

Plotting by hand should be the first step in analysis and need not be time-consuming. Plot the effect for a given logarithm of concentration, whether the effect is size attained, percent inhibition of reproduction, or some other quantitative effect (see sample results in Section 10, Figures 22–31, and Appendix P, Figure P.1).

A list of advantages of the rough plot follows.

(a) The plot will show anything unusual in the results. It might be something of considerable interest biologically, which would not be noticed otherwise.

(b) The general shape or form of the dose-effect relationship will be evident, which could prevent forcing the data into a mathematical model that was unsuitable.

(c) A rough estimate of the endpoint can usually be made from the graph. If the endpoint from mathematical analysis does not agree reasonably, the cause of the divergence should be sought. Sometimes, that could help the investigator to avoid reporting a result which contained an unexpected arithmetic oversight or an incorrect transfer of data [45].

Alternatives could be plotting the raw data onto a computer-generated graph, or vice-versa.

6.3.2 Choosing the Method

Linear/nonlinear regression analysis is the method of choice for quantitative sublethal tests in Canadian laboratories of environmental toxicology. Guidance on these methods is available (Section 6.5.8), and Environment Canada now requires regression analysis as the primary choice for tests of growth and reproduction in soil organisms (EC, 2004a–c).

The most common method used in the past, and the easiest choice, has been *Smoothing and Interpolation* (see Section 6.4). The notable deficiencies of this

[45] Good computer programs often include a component to plot the results, which is useful, but as pointed out in Section 4.2.2, it is not a substitute for a graph drawn by hand. If there had been an error in entering data, the computer-drawn graph and the calculations would agree, but both would be incorrect.

method have been known to Canadian investigators for many years and were described at the Quebec City meeting of the Statistical Advisory Committee (Miller *et al.*, 1993; see Section 6.4.1). The participants recommended regression analysis as an alternative procedure and expressed the need for guidance in selecting the appropriate model. As mentioned, this guidance is now available.

6.4 Smoothing and Interpolation

Key Guidance

- *This method has been in standard use across North America, and should now be phased out in Canada in favour of regression methods. It is convenient because the only assumption about the pattern of results is that effect increases with concentration.*
- *The ICp is estimated by interpolation between the two adjacent data-points, which is less desirable than a regression approach using all the data.*
- *An initial step in analysis adjusts the raw data so they are monotonic, which to a limited extent, can make use of the wider distribution of data.*
- *Calculation of the ICp is simple enough to do by hand, but a convenient computer program is freely available ("ICPIN").*
- *At present, the computer program fails to adopt a logarithmic scale of concentration. Canadian users of the program must enter the concentrations as logarithms.*
- *Confidence limits cannot be calculated by the usual methods. Instead, the computer program is used for a "bootstrap" estimate. The computer resamples from the original measurements at least 240 times (a recommended minimum), to estimate confidence limits.*

6.4.1 General Critique

This method of interpolation was introduced by the USEPA (Norberg-King, 1993) and is available as the computer program ICPIN. Estimates by linear interpolation have some conceptual problems (see following text), but the interpolation method is versatile. It has been the usual way of obtaining a quantitative point estimate in North America, in the previous absence of a convenient statistical package for regression. ICPIN is little known in Europe at the time of writing (Niels Nyholm, 2001, personal communication, Technical University of Denmark, Lyngby, Denmark).

Some general deficiencies of smoothing and interpolation have been listed as follows, by the Statistical Advisory Group (Miller *et al.*, 1993).

- inefficient use of data, since the method interpolates between only two concentrations bracketing the endpoint, and neglects the overall relationship between effect and concentration (except for some general influence of smoothing);
- sensitive to any irregularities or peculiarities of the two concentrations used;
- fails to use logarithm of concentration, introducing a slight bias into the calculation of ICp, towards a higher value; and,
- the "bootstrap" method of calculating confidence limits sometimes produces unrealistically narrow limits.

The three assumptions that follow are implicit in the Smoothing and Interpolation method. (Sometimes this is called an "assumption-less" method because it does not postulate any particular form of dose-effect curve, but nevertheless there are some assumptions.)

- The effects must increase monotonically at each successive higher concentration (or at least, they should not decrease). If this requirement is not met, it is imposed by mathematical manipulation.
- The method assumes that effects increase linearly *between* two successive concentrations. (Sometimes described as following a "piece-wise" linear function, a term with various connotations.)

- The effects should come from a representative sample of test data that is random and independent, an assumption that applies to most statistical analyses.

In practice, problems with the required assumptions are seldom recognized when carrying out this procedure. For the first requirement (monotonic series), the data are simply adjusted as necessary, to make the series monotonic. There is no way of testing the second assumption (piece-wise linearity). There is seldom, if ever, a test of the third assumption (random, independent results).

The investigator has little opportunity to ascertain whether the method is producing a reliable result. The method should be used with caution if the effects diverge strongly from monotonic. The method is particularly inappropriate for data showing hormesis (Section 10.3), such as some tests with the alga *Pseudokirchneriella subcapitata*. The method would also be risky if successive concentrations caused very low and very high effects (USEPA, 1995). Nevertheless, the smoothing procedure will hide such irregularities, and the method is often used for irregular data. Caution in such cases would mean subjective comparison with the original data and the hand-drawn graph.

6.4.2 Steps in Analysis

A generalized description of the mathematical procedures for estimating an IC25 by smoothing and interpolation is provided in the following steps, because the method has been so widely used. When an example is needed, it is the weight of fish at the end of a toxicity test. Appendix N provides a very detailed description of the analysis, and the computer program ICPIN. Investigators who use the method would do well to understand the steps in Appendix N.

(1) As a subjective check on the quality of data, plot the unadjusted average weight of each group of fish against the logarithm of concentration.

(2) Start the linear interpolation by smoothing the data if the average weight increases between one concentration and the next higher concentration.

The IC25 is estimated by a simple linear interpolation between the two concentrations which bracket it. Hand calculations (see next steps), follow the same steps as the computer program ICPIN. The steps sound complicated but are actually rather simple arithmetic.

(3) Calculate the weight that represents the endpoint. It is 75% of the average weight of the control fish, i.e., a 25% reduction.

(4) From the result of step (3), subtract the average weight at the concentration immediately below the IC25. This will normally be a negative number, in a growth experiment.

(5) From the average weight at the concentration immediately above the IC25, substract the weight at the concentration immediately below the IC25. This is also likely to be a negative number in a growth experiment.

(6) Divide the result of step (4) by the result of step (5).

(7) Calculate the difference between the logarithms of the two concentrations below and above the IC25. Subtract the logarithm of the concentration just below the IC25, from the logarithm of the concentration above the IC25.

(8) Multiply the result of step (6) by the result of step (7). This is the upward movement of concentration to the IC25, from the concentration immediately below it.

(9) Add the result of step (8) to the logarithmic concentration immediately below the IC25. The result is the IC25 as a logarithm.

The ICp cannot be estimated if there is not one test concentration lower than the ICp, and another higher. One can only say that the ICp is lower than the lowest concentration tested, or greater than the highest concentration, as the case may be.

A computer is necessary for estimating the confidence limits (see Section 6.4.3 and Appendix N).

6.4.3 The Computer Program ICPIN

ICPIN runs on personal computers and is available within commercial packages; however, free copies are widely available (Appendix N). ICPIN is easy to use, has a very clearly written set of instructions, and the steps for entering and manipulating data are obvious.

ICPIN carries out all steps (1) to (9) listed in Section 6.4.2, before it calculates the confidence limits. However, *the investigator must manually determine the logarithms of test concentrations and enter those logarithms*, rather than arithmetic concentrations as specified in the instructions of the program.

A computer must be used to obtain 95% confidence limits about the ICp. A technique called "bootstrapping"must be applied, because the usual statistical methods cannot be used after interpolation. ICPIN does this by calculating a series of ICps that might have been obtained, based on resamplings of the original observations (see Appendix N). To do this, the toxicity test must have replicates. From the distribution of the hypothetical ICps, it is possible to calculate confidence limits for the estimated ICp.

6.5 Point Estimates by Regression

Key Guidance

- *Regression techniques represent the method of choice for estimating ICp. There are many relevant publications, and specific guidance has recently been made in Canadian test methods.*
- *Most patterns of sublethal quantitative effects can be fitted by nonlinear regression. Potential patterns are reviewed here, with step-by-step guidance in Appendix O.*
- *No single model of nonlinear regression can be suitable for all the dose-effect patterns encountered. Investigators can fit most patterns by selecting from five defined models. Subsequently, the data can be fitted and the endpoint estimated by applying programs from a general-purpose statistical package. The choice and subsequent analysis require some statistical knowledge, and not simply the rote application of a computer program. At least one "off-the-shelf" package of statistical software for environmental toxicology offers a wide selection of models for nonlinear regression.*
- *Nonlinear regression can fit effect-distributions showing hormesis. One suggested model allows for the hormesis, yet uses the true control for deriving the endpoint.*
- *Environment Canada now requires regression analyses in the recent toxicity test methods for assessing growth and reproduction in soil organisms.*

This subsection gives some background on using regression, then proceeds towards specific procedures for the regression methods now required in most Environment Canada tests for soil toxicity.

There has been widespread interest in improving methods of analysis for sublethal toxicity tests, which has been largely directed towards regression techniques. A workshop on statistical methods sponsored by SETAC-Europe in 1995 was attended by some two dozen participants from many countries (Chapman *et al.*, 1996a), and a similar workshop was sponsored by the OECD (Chapman, 1996).

Regression analysis has been a tool in statistics for a long time (Draper and Smith, 1981), and likely provides the best method for estimating quantitative sublethal endpoints. There has been appreciable development and standardization of regression methods for environmental toxicology.

Regression or regression analysis is a mathematical description of the relationship between two or more variables. In this document, the *dependent variable* is the observed effect. Its value depends upon the *independent variable*, the concentration, or perhaps on more than one variable if there are modifying conditions. The data are fitted mathematically to a selected model, then (in toxicology) an endpoint is picked from the model. Standard mathematical techniques can describe a regression to convey useful information to others. Effects at high and low concentrations can be predicted, and confidence bands can be estimated. The selected model should conform to the pattern of the data, even if that model has no particular biological basis or little theoretical rationale (Moore, 1996).

A regression approach has the problem that no single model will fit the diverse dose-effect curves from sublethal tests. A spectrum of models is required, with guidance on choosing the appropriate one.

Transformation of effects-data is a possible approach for making the results fit a relatively simple linear model. There are advantages and disadvantages (see Section 2.9.1), and generally it is preferable to avoid transformations.

6.5.1 ABCs of Regression

The common requirements and essential steps in any regression (in the context of toxicity tests) can be stated simply as follows.

(1) Compile the data-set.
The test has a fixed set of values for the *independent variable* (concentration). At each of those values, observations are made for the *dependent variable* (effect).

(2) Choose a model.
Some relationship between the dependent and independent variables is proposed by the investigator. It is specified as a mathematical function such as a straight line or logistic curve.

(3) Select a procedure for fitting the relationship to the data.
First, check that assumptions of the model are met (e.g., normality of data). Then, the parameters of the model are usually estimated by minimizing the squared deviations of the observations from the curve which is serving as the model. Environment Canada's standard method is described in Section 6.5.8.

(4) Calculate and consider the goodness of fit of the data to the model.

(5) Carry out the inverse estimation of the concentration predicted to cause the selected degree of effect (the endpoint, e.g., the IC25).

(6) Find the confidence limits on the endpoint, also by inverse estimation.

The calculations are normally carried out using a computer program for regression.

6.5.2 Concepts: Linear, Nonlinear, GLM, and GLIM

> *...all models are wrong; some, though, are more useful than others and we should seek those.* (McCullagh and Nelder, 1989)

Key Guidance

- *In linear regression, the "linear" refers to the relatively simple nature of the equation. The parameters (a, b, etc.) can be estimated by evaluating a single formula.*
- *In nonlinear regression, the parameters are not independent of other parameters. An iterative approach is required to estimate the model parameters.*
- *General linear model (GLM) describes a class of similar models. The class includes simple linear regression, analysis of variance, analysis of covariance, repeated measures, and others. Generalized linear model (GLIM, GLiM) is a broader version of the approach used for the GLM. Statisticians use it to estimate the parameters of models which include exponential, binomial, logistic, Poisson, and log-normal distributions. The concept is quite advanced and as yet is not widely used in environmental toxicology.*
- *The advantage of nonlinear regression is that all of the data are used for a point estimate with confidence limits, for various shapes of effect-curves including hormesis. The control measurement is included in the regression. Some knowledge and judgement must be applied, however, in selecting the model and applying the statistical procedures.*

Non-mathematicians must keep in mind that when statisticians refer to *linear* or *nonlinear*, they are not describing the shape of a mark drawn by a pencil on a piece of paper. They are referring to the relationships of elements within an equation. In statistical parlance, an expression is linear in its parameter(s) if a solution for the parameter can be written down that refers only to the data and not to another parameter. Linearity refers to the relationship between the effect and the *parameters* of the model, not the relationship between the effect and the independent *variable* or variables. The simplest example of a linear model is Equation 4 for a straight line (Section 6.5.3). The values of α and β can be determined by means of a

little arithmetic, based on the observed values of X and Y. Beyond that, a quadratic equation (Equation 5) is still a linear model because its parameters can also be estimated from the observed data.

In *nonlinear regressions*, it is not possible to estimate the values of the parameters from the observed data, in one step. An iterative approach must be used to solve the equations which estimate each parameter. (See Section 6.5.4.)

At a more complex level, the words "linear models" are used in two confusingly similar terms which differ in meanings to statisticians. Both are relevant to analysis of toxicity tests. The first term, *General linear models (GLM)*, represents a general class of models with a single dependent variable (Section 6.5.10). The class includes familiar models such as ANOVA and regression and also more complicated models such as ANCOVA and repeated measures. General linear models apply only if the data (such as weight of organisms) are normally distributed. Under the strict definition, binomially distributed data such as mortality, would not be included.

The second term is *Generalized linear models (GLIM or GLiM)*, representing an even larger category of models that includes GLMs as a sub-category. GLIMs allow an investigator even more scope for analyzing quantal or quantitative effects that arise from either simple or complex arrangements of independent variables in an experiment (Section 6.5.11).

6.5.3 Linear Regression

The familiar relationship of a straight line (Equation 4) represents a *linear model*. (Curved lines can also be included in the category.)

$$Y = \alpha + \beta X \qquad [\ \textit{Equation 4}\ ]$$

The formula describes the relationship between a measured effect *Y* which is the *dependent variable*, and a predictor *X* which is the *independent variable* and in this case would likely be the logarithm of concentration.

In Equation 4, alpha and beta (α and β) are *parameters*. Alpha is the intercept of the straight line with the y-axis, i.e., the value of the dependent variable (Y) when the independent variable (X) is zero. Beta is the slope of the regression, i.e., the rise in Y for unit increase of X. For a given set of data, the parameters would be estimated by some mathematical means. Often this is done by *least squares*, which estimates the values of the parameters that minimize the sum of squares of deviations of observed values from the model.

The relationship might be causal as implied by the term dependent", or it might only be a correlation. There are only two variables being considered, so this is *simple regression*, hence the term *simple linear regression* (Zar, 1999). (There might be more than one independent variable, in which case a more complex formula would describe it, as indicated below under "Multiple regression".)

Toxicity tests might sometimes yield a relationship between effect and concentration that appears straight, at least for the central part of the regression. Calculating the best fit of the line (model) to the data could then be done by conventional means for dealing with linear regressions, such as least squares. Indeed, simple regressions have been used to describe results, especially for sublethal effects such as growth (e.g., Rowe *et al.*, 1983).

Linear regression is a simple model. If a set of toxicological data fits the regression satisfactorily, then the relationship can be used in a predictive fashion. For any given or selected value of X (e.g., log concentration), one can calculate from the equation, the predicted value of Y (say the weight of fish exposed to that concentration). It is important that the values of the independent variable X are created and measured without error. [As described in Sections 6.2.4 and 9.4), the toxicological investigator finally makes an inversion to estimate the concentration (and its confidence limits) that can be expected to cause a selected level of effect (e.g., 25% reduction from control performance, the ICp.)]

A more complete, correct, and explicit description of the model would add subscripts to Equation 4. Although they will usually be neglected in the present document, they are implicit, and investigators should expect to encounter them elsewhere. Subscripts would be required if an equation represented a set of observations in a test. Subscript "I" would indicate each of the organisms or measurements in the test, and subscript "j" would indicate levels of toxicant.

The modification of Equation 4 would be Equation 4a.

$$Y_{ij} = \alpha + \beta X_j \qquad [Equation\ 4a]$$

Because data points would show scatter about the fitted line, an error term (ϵ_{ij} or e_{ij}) is added to the equation. The "e" represents random variability of an individual measurement "I" at the "jth" concentration. The complete *linear regression* is Equation 4b.

$$Y_{ij} = \alpha + \beta X_j + e_{ij} \qquad [\ Equation\ 4b\ ]$$

Multiple regressions. These are included in the category of linear regression. The name indicates that the dependent variable is governed by two or more independent variables (X_1 and X_2 in Equation 5). For example, the toxicity of a metal might depend not only on concentration of the metal but also on temperature of the medium. Equation 5 might represent a regression with four parameters α, β_1, β_2, and the variance (sigma).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2 \qquad [\ Equation\ 5\ ]$$

Equation 5 falls into several categories. It can be called a *multiple regression* because it has several terms. It is also a *quadratic function*, because of the added final, so-called quadratic term. Statisticians point out that in using such a model for toxicity data, it should be limited to describing a *local* effect. In theory, the quadratic is inappropriate because it will predict a decreased effect at some high concentration, for the usual dose-effect situation. As stated, however, it can be useful for describing local effects within a limited range of concentrations.

6.5.4 General Aspects of Nonlinear Regressions

The linear relationship (Section 6.5.2) is relatively simple, and is often insufficient for describing a complex relationship of effect with concentration. The investigator would have to select a more complex *model* (i.e., a mathematical function), in an attempt to fit the toxicity data, and the shape of the dose-effect relationship might well lead to a nonlinear regression model. Two or more of the model's parameters might be functions of each other in multiplicative fashion, as in exponential growth, shown by Equation 6 (Zar, 1999). Clearly, estimating the parameters of such an equation will be more complicated than for a linear regression.

$$Y = \alpha\, \beta^X \qquad [\ Equation\ 6\ ]$$

Often, a function describing a sigmoid shape would be appropriate in environmental toxicology. Two nonlinear models that have often been found suitable, are the logit and Weibull equations. The logit model is symmetrical while the Weibull model is asymmetric (see Sections 4.5.1, 4.5.2, and Appendix J). Other useful models are detailed in Section 6.5.8.

Once the function (model) is specified, "best" estimates of its parameters are found by maximum likelihood or least squares techniques. As mentioned previously, an iterative approach must be used to solve the equations which estimate each parameter.

The iteration used with nonlinear regression might be described informally as initial "guesses" of the values of parameters in the model, guesses made by the investigator or the program used. In successive iterations, these initial values are changed upwards or downwards by the program, to approach more closely to a fit of the observed data. In other words, the program searches for an optimum value for each parameter. This model can be visualized as a group of small hills for the various parameters, with each parameter having an optimum value at the top of a hill. The program can determine at the time of each iteration, the slope of the hill locally, and hence the proper direction to go in the next iteration, to move towards the optimum value of the parameter ("top of the hill"). When estimates for all parameters remain essentially constant in successive iterations, the procedure has *converged* to a final solution, i.e., it has achieved the best estimates of the model parameters for the particular set of data.

The OECD (2004) points out that it can be important to make realistic initial "guesses" for the parameter values. The final estimates might depend on that initial choice, because there might be several local maxima or optima for a given parameter. In the visualization, it might be thought of as several small sub-peaks scattered on the slopes of a big hill. Since the program, in any one iteration, can appreciate only the slope in the immediate vicinity, not the complete shape of the hill, it might work to the top of a sub-peak and remain there, in an undesirable "convergence". Hence the importance of an initial realistic starting point, near the main peak.

One method of arriving at the final estimates of variables in an equation is by using *least squares* techniques. Iteration with least squares was referred to in solving probit regression for quantal data (Section 4.5.3). In least squares, the predicted and observed values of the dependent variable (toxic effect) are compared at given levels of the independent variable (log concentration). The difference between predicted and observed is called a *residual*, and residuals become smaller as a line fits better. Residuals are squared and summed and that "sum of squares" is taken as an assessment of fit. Obviously, the lowest sum of squares indicates the best fit, giving rise to the term "least squares".

The least squares solution for parameters of an equation will often be equivalent to those obtained by *maximum likelihood solution*, a more sophisticated, complex, and mysterious mathematical approach.

As indicated in subsections 6.5.7 and 6.5.8, use of nonlinear regression requires some judgement and familiarity with the mathematical techniques. A general statistical software program is often used, although there is at least one statistical package designed specifically for environmental toxicology (CETIS), that provides a wide choice of models. The general methods for regression have existed for some time in standard statistical texts and packages, but it has often been a time-consuming exercise for toxicologists to develop their own expertise (Moore, 1996). The useful techniques in toxicology have been described by Newman (1995).

Guidance on nonlinear regression is available in textbooks such as Bates and Watts (1988). An investigator starting to use nonlinear regression would benefit from the advice of an experienced statistician (Section 2.1). A naive investigator might produce incorrect results by failing to satisfy the assumptions of a technique, choosing an inappropriate model, etc. Further information is given in Sections 6.5.7 to 6.5.9.

Advantages of Nonlinear Regression. For analysis of toxicity data, regression is much more defensible than smoothing and interpolation, or hypothesis testing. The test data will dictate the type of regression. If a linear regression fits, it should be used; if it is not suitable, then a nonlinear model becomes a better choice. Some of the general advantages of regression, and particular advantages of nonlinear regression, can be listed as follows:

- all of the data from the test are used;
- a point estimate is obtained, the ICp;
- confidence limits of the ICp are obtained;
- any value of p can be used, e.g., IC25; various shapes of concentration-effect curves can be accommodated;
- the control measurement is included in the fitted regression; and
- hormesis can be accommodated without compromising the control effect.

The main disadvantage is that there cannot be a simplistic "black-box" computer program designed for toxicology. The investigator must exercise some knowledge and judgement in selecting the model and applying the statistical procedures.

6.5.5 Choosing a Regression Model

Key Guidance

- *It is prudent to select a model that is adequate, but as simple as possible. A "parsimony of parameters" is desirable -- each one added to the model loses a degree of freedom.*
- *One way to keep simplicity in the model is to eliminate parameters which are correlated to another parameter which is already in the model.*
- *The model might fail to fit because of a poor choice of model, an overly complicated one, outliers, or mistakes in coding. Sometimes the original data might not cover the upper or lower range of the model.*
- *Plots of residuals against the predicted values provide a visual assessment of the fit of a model, and a visual check of the data is always needed. For linear regression, the coefficient of determination ("R^2 value") can be used to assess fit.*

In choosing a model, investigators must consider their own priorities as well as technical matters. The "best" model for one person might be that with the least error of prediction, while another person might emphasize a parsimonious function, or another that shed the most light on biological mechanisms. Some aspects in the choice of model are mentioned in the following text.

It is prudent to adopt relatively simple but adequate models, and avoid overly complex models. Certainly a polynomial equation with enough terms could be made to fit almost any unusual pattern of effect, but adding additional parameters has penalties, such as loss of degrees of freedom, and widening confidence limits. In their text on generalized linear models, McCullagh and Nelder (1989) caution against using many parameters to get a close fit to data. "In so doing, however, we have achieved no reduction in complexity ... simplicity, represented by parsimony of parameters, is also a desirable feature of any model; we do not include parameters that we do not need. Not only does a parsimonious model enable the research worker ... to think about his data, but one that is substantially correct gives better predictions than one that includes unnecessary extra parameters." In the preceding quotation, "unnecessary" might be interpreted as a parameter that was not statistically significant. Another expert points to the possibility that an obscure biological interpretation might arise from a complex model with a four-parameter equation: "a fit may look smart, but how to use the results when the computer people have gone home?" (Nyholm, 2001).

One example of unnecessary complexity would be for observations that were highly correlated (e.g., length and weight of organisms), and investigators should be wary of using parameters for each in a regression model. That can lead to a problem of "multicollinearity" and error messages or failure of the fit. Statistical packages usually produce a correlation matrix for the parameters and it should be examined; high correlations might indicate that one of a pair of variables could be omitted.

In multiple regressions, it is possible to test whether all the variables are necessary, and this is highly recommended. The preferred way is to do a series of fits with and without parameters of interest, and compare results (see Section 6.5.6 under the heading "Explained variability in regression"). Another method that is mentioned in some statistical texts, but is not recommended here, is checking each parameter with a t-test (sometimes provided in the statistical package). The null hypothesis would be that the parameter equalled zero, and if the t-test did not disprove that, the parameter would be deleted from the regression.

Weighting might be required, as explained by Nyholm *et al.* (1992): "If the variance of the data points is constant (constant absolute error), nonlinear regression can be carried out directly with no weighting, ... otherwise a proper statistical weighting must be used. The weighting factors should be inversely proportional to the variance of the data points ...". This requires replication and testing for equal variances as described in Section 6.5.8.

6.5.6 Adequacy and Fit

Failure to fit. The model might "fail to converge" in which case there would be no fit and no estimate of the parameters. One reason for failure to converge can be multicollinearity (see Section 6.5.5). Sometimes, satisfactory estimates of the parameters might not be obtained, even after satisfactory convergence. Some possible reasons can be listed.

- **Poor choice of model**. An unsuitable model cannot be expected to fit.

- **Outliers**. Even one outlier could prevent convergence. The outlier must not be arbitrarily omitted in a modelling process, but instead given objective consideration by methods such as those mentioned in Sections 6.5.8 and 10.2.

- **Mistakes in coding**. Inaccuracy or errors in coding (see Glossary) can produce nonsensical results.

- **Range of data**. Procedures might be satisfactory, but the original data could be deficient. Values might fail to cover the upper or lower range of the model. Quantal data should span the range from no effect to complete effect. Quantitative data should be represented in each section of the model's shape. This is a relatively common deficiency, discussed in Section 2.2. Range-finding tests can remedy this problem.

- **Overly complicated model**. (e.g., multicollinearity, Section 6.5.5). A simpler model should be adopted if observations fail to cover part of the desired distribution.

When parameters have been successfully estimated, the investigator must decide if the model gives an adequate description of the variability. Most statistical packages provide an F-test; if that test produces a p-value that is less than 0.05, it can be concluded that the regression model describes a significant proportion of the data, with 95% confidence. Other assessment should continue as described in the following text.

Explained variability in regression. *Plots of residuals* against the predicted values provide a visual assessment of satisfactory fit of a model (see Glossary for residuals). Certain problems might be revealed. A series of residuals above or below predicted values could indicate inadequate fit or correlated observations. A vee shape in residuals indicates heterogeneity of variance. A divergent pattern suggests an incorrect model (see Section 6.5.8 and Appendix O).

Other common sense appraisals should follow the plot. Was the range of tested concentrations wide enough to show the scope of effects? Does a plot of the fitted regression do a reasonable job of representing the actual observations? Does the shape of the model fit with mechanisms that are thought to govern the effect? Have outliers influenced the fit to an undue extent? If negative answers predominate, the investigator would be wise to consult a statistician.

For linear regression, the *coefficient of determination* or R^2 ("the R^2 value") is the sum of squares explained by fitting the model ($SS_{regression}$), divided by the total sum of squares (SS_{total}) about the mean. Values are often expressed as a percentage and could theoretically range from zero (nothing explained) to 100% (a perfect fit for the model). The 100% will not be encountered, and very high results are not necessarily desirable. Such high results suggest a complex model with many parameters and the associated drawbacks (see Section 6.5.5). The coefficient of determination cannot be applied to nonlinear models.

The OECD (2004) cautions against blindly applying a statistical test of goodness of fit in a strict and absolute manner (i.e., the model either fits or does not fit). The OECD guidance document states that "A visual check of the data is always needed and may overrule a goodness-of-fit test." That advice is intended to encourage the investigator to check that the data provide sufficient information to confine the model. For example, if there had been additional data for intermediate levels of dose, could that have changed the shape of the relationship? The OECD also points out that data with only a few test treatments can more easily pass a goodness-of-fit test. Alternatively, a good set of data with a single deviant treatment/effect could result in rejection of a model that, otherwise, followed the data perfectly.

There are other means of evaluating fit. ANOVA can summarize a regression model, and the overall F-test checks the null hypothesis of adequate fit. Another form of R^2 would use only the denominator to describe residual error. A low value is desirable, but again, over-parameterization can be one cause of low error. A superior version of R^2 is provided by Mallows (1973), whose C_p penalizes models that are over-parameterized. Similar measurements that should be recognized as superior if encountered are the Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC).

6.5.7 A Recent Example of Nonlinear Regressions

Key Guidance

- *A group of Canadian authors developed procedures that have served as the basis for Environment Canada's standard approach of using regression for point estimates of quantitative sublethal endpoints.*
- *These authors applied the linear and nonlinear regression models, which were available in a standard statistical package (SYSTAT), to their tests of soil toxicity to plants.*
- *The authors found that most sets of results could be fitted satisfactorily by one of five models: linear, logistic, logistic with hormesis, exponential, and Gompertz.*

Clear illustrations of fitting nonlinear regressions to sublethal quantitative data were provided by Stephenson *et al.* (2000), with explanation of the same research by Koper (1999). These investigators obtained useful estimates of the sublethal toxicity of contaminated soils to several species of plants. Their methods have been further developed by Environment Canada as required procedures in new soil tests (EC, 2004a–c; see Section 6.5.8).

Stephenson *et al.* (2000) illustrated the general form and explained equations for three models, the **logistic**, **logistic with hormesis**, and **exponential**. Other useful models were added to their nonlinear package (Koper, 1999). One was the sigmoidal **Gompertz**, another was a standard **linear equation**. A parameter was added to the **exponential** model to allow the asymptote to be a value other than zero.

The techniques and difficulties of nonlinear regression are described briefly by Stephenson *et al.* (2000) who provide a flow sheet, similar to Figure 16, as a guide to the steps in selecting the most appropriate model. An initial estimate had to be made by the investigator, for each parameter in the model. (Realistic initial estimates must be made, or an anomalous endpoint could be chosen by the statistical program, see Appendix O.) The parameters of the fitted equation were then estimated using iterative calculations. Stephenson *et al.* (2000) noted that having too many parameters could prevent estimates from being made. A suitable strategy was to use the simplest suitable model (Section 6.5.5), enough replicates, and up to 12 treatments. The need for equal variances in the treatments was also troublesome, because inequalities might lead to an inflated estimate of the standard error and confidence limits. To correct for this, observations were weighted using the inverse of the variance for observations at each treatment (see Section 2.6). Good estimates of the variance were required to do this, sometimes $\geq$9 replicates per concentration. Koper (1999) recommended that if weighting was necessary, calculations should be done for both weighted and unweighted distributions, then the results and distribution of residuals can be compared.

Koper (1999) pointed out that through computers, nonlinear regression has become feasible for routine use in laboratories. Models were reparameterized so that the calculations automatically generated the estimate of ICp and its confidence limits (see Section 6.5.12 on reparameterization). The process for reparameterizing was developed from the methods of Van Ewijk and Hoekstra (1993) and Hoekstra and Van Ewijk (1993). The analyses were run using the statistical software package SYSTAT 7.0.1.

Problems with fitting could be caused by collinearity, which occurs when parameters are highly correlated, or when a value close to zero in the denominator of a matrix was inverted as part of the calculations. Other items of possible statistical difficulty were convergence, choosing a maximization algorithm, local versus global maxima, and comparing nested and non-nested models.

The procedures of Stephenson *et al.* (2000) had some requirements. Data must bracket the ICp (which would also be beneficial or essential for other methods). At least 10 or 12 treatments were recommended, to show the shape of the relationship and choose the model. The large number of treatments also contributed to the success of the computer calculations. The number of replicates per treatment could be two, although these investigators had up to six replicates. There was no need to have the same number of replicates at each concentration.

Readers emulating this work should be aware that Stephenson *et al.* (2000) failed to use logarithms of concentration. Investigators should use log concentration in the scatter diagrams and in the calculations, as in the standard procedure of Environment Canada (Section 6.5.8). As explained in Section 2.3, this is a question of proper scientific procedure, not just of statistical procedure or whether the model is capable of handling arithmetic values of concentrations.

6.5.8 Environment Canada's Method for Regression Analysis

Key Guidance

- *The new tests of Environment Canada for soil toxicity require linear/nonlinear regression as first choice for estimating the ICp. Specific procedures are outlined for SYSTAT or other statistical packages.*

- *Investigators choose from five models: linear, logistic, Gompertz, exponential, and logistic adapted for hormesis. The models have been re-parameterized to directly estimate the ICp and confidence limits.*
- *Assumptions of normality and homoscedasticity of the residuals must be satisfied before the estimate is made.*
- *If regression methods are not successful, then the ICp should be estimated by interpolation using the program ICPIN.*

The new *Biological Test Methods* published by Environment Canada for earthworms, plants, and springtails (EC, 2004a–c) require linear and nonlinear regression to be applied as the primary method of analysis for the quantitative sublethal data. Only if results are unsuitable for regression, is the investigator allowed to fall back on less desirable methods of analysis.

Upon completion of a multi-concentration test, the ICp and 95% confidence limits must be calculated using one or more of a series of linear and nonlinear regression models proposed by Stephenson *et al*. (2000). The models have been re-parameterized using techniques from van Ewijk and Hoekstra (1993), to automatically generate the ICp and its 95% confidence limits for any specified value of 'p' (e.g., IC25 or IC50). The models include one linear model, and the following four nonlinear regression models: exponential, Gompertz, logistic, and logistic adjusted to accommodate hormesis[46]. Instruction is provided in Appendix O for applying linear and nonlinear regression using Version 11.0 of the statistical program SYSTAT[47]. However, any statistical software capable of linear and nonlinear regression may be used (see the end of this subsection for comments on other statistical software).

Descriptions of the five models follow, with further information provided in Appendix O. The exponential model given below is a general version, while the coded version in Appendix O has some specific modification.

Exponential model:

$$Y = a \times (1 - p)^{(C \div ICp)}$$

where:

Y	=	dependent variable (e.g., number of juveniles, root/shoot length, or dry mass)
a	=	the y-intercept (i.e., the control response)
p	=	desired value for 'p' (e.g., 0.25 for a 25% inhibition)
C	=	the test concentration as a logarithm
ICp	=	the ICp for the data-set

Gompertz model:

$$Y = t \times \exp[\log(1-p) \times (C \div ICp)^b]$$

where:

Y	=	dependent variable (e.g., number of juveniles, root/shoot length, or dry mass)
t	=	the y-intercept (i.e., the control response)
exp	=	the exponent of the base of the natural logarithm
p	=	desired value for 'p' (e.g., 0.25 for a 25% inhibition)
C	=	the test concentration as a logarithm
ICp	=	the ICp for the data-set
b	=	a scale parameter, estimated to be between 1 and 4, that defines the shape of the equation

---

[46] A hormetic type of response (low-dose stimulation) might be found in sublethal observations at the lowest concentration(s), i.e., performance at such concentration(s) is enhanced relative to that in the control. For instance, there might be more progeny produced in low concentrations than in the control, or the weights of individuals might be higher than in the control. This response is a real biological phenomenon, not a flaw in the testing. Such data should be analyzed using the hormesis model. The hormetic effects are included in the regression, but do not bias the estimate of the ICp. An estimated IC25 would still represent a 25% reduction in performance from that for the control.

[47] The latest (Version 11.0 or later) version of SYSTAT™ is available for purchase by contacting SYSTAT Software, Inc., 501 Canal Boulevard, Suite C, Point Richmond, Calif. 94804-2028, USA, phone: 800-797-7401; www.systat.com/products/Systat/.

Hormesis model:

$$Y = t \times [1 + (h \times C)] \div \{ 1 + [(p + (h \times C)) \div (1 - p)] \times (C \div ICp)^b \}$$

where:

Y	=	dependent variable (e.g., number of juveniles, root/shoot length, or dry mass)
t	=	the y-intercept (i.e, the control response)
h	=	describes the hormetic effect (estimated to be small, usually between 0.1 and 1)
C	=	the test concentration as a logarithm
p	=	desired value for 'p' (e.g., 0.25 for a 25% inhibition)
ICp	=	the ICp for the data-set
b	=	a scale parameter, estimated to be between 1 and 4, that defines the shape of the equation

Linear model:

$$Y = [(-b \times p) \div ICp] \times C + b$$

where:

Y	=	dependent variable (e.g., number of juveniles, root/shoot length, or dry mass)
b	=	the y-intercept (i.e., the control response)
p	=	desired value for 'p' (e.g., 0.25 for a 25% inhibition)
ICp	=	the ICp for the data-set
C	=	the test concentration as a logarithm

Logistic model:

$$Y = t \div \{1 + [p \div (1 - p)] \times (C \div ICp)^b \}$$

where:

Y	=	dependent variable (e.g., number of juveniles, root/shoot length, or dry mass)
t	=	the y-intercept (i.e, the control response)
p	=	desired value for 'p' (e.g., 0.25 for a 25% inhibition)
C	=	the test concentration as a logarithm
ICp	=	the ICp for the data-set
b	=	a scale parameter (estimated to be between 1 and 4) that defines the shape of the equation

The general process for selecting the most appropriate regression model, and subsequent statistical analysis for quantitative toxicity data, is outlined in Figure 16. The selection process begins with an examination of a scatter plot or line graph of the test data to determine the shape of the concentration-response curve. Its shape is then compared to available models so that one or more model(s) that best suit(s) the data is (are) selected for further examination (see Figure O.1, Appendix O, for examples of the five models).

Once the appropriate model(s) is (are) selected for further consideration, assumptions of normality and homoscedasticity of the residuals are assessed. If the regression procedure for one or more of the examined models meets the assumptions, the data (and regression) are examined for the presence of outliers. If there is an outlier, the test records and experimental conditions should be scrutinized for human error. Then the analysis should be performed with and without the outlier(s), to examine the effect of the outlier(s) on the regression. A decision must be made on whether or not to remove the outlier(s) from the final analysis, considering natural biological variation, and other biological reasons that might have caused the apparent anomaly. Additional guidance on the presence of outliers and unusual observations is provided in Section O.2.4 in Appendix O as well as in Section 10.2. Additional guidance from a statistician familiar with dealing with outliers is also advised.

If there are no outliers present or none are removed from the final analysis, the model that demonstrates the smallest residual mean square error is selected as the model of best choice.

Normality should be assessed using the *Shapiro-Wilk's test* as described in Appendix P, Sections P.2.1 and P.2.2. A normal probability plot of the residuals may also be used during the regression procedure, but is not recommended as a stand-alone test for normality, because detection of a 'normal' or 'non-normal' distribution would depend on the subjective assessment by the user. If the data are not normally distributed, then the user is advised to try another model, consult a statistician for further guidance, or to analyze the data by the less-desirable method of linear interpolation

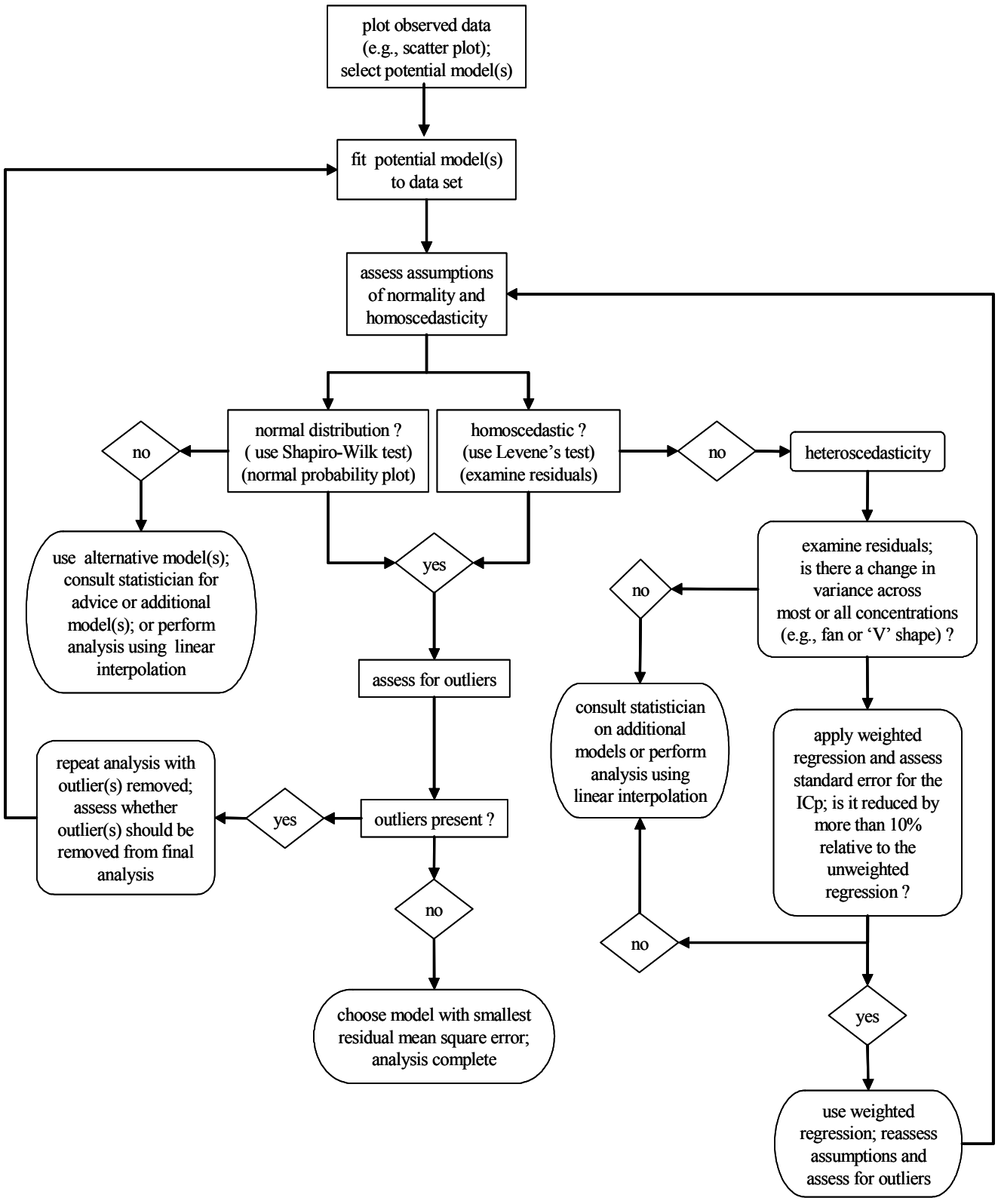


Figure 16 The general process for selecting the most appropriate model and completing the statistical analysis for data on quantitative toxicity (adapted and modified from Stephenson *et al*. 2000).

using ICPIN (see Section 6.4 and Appendix N). In recent EC soil tests, the ICPIN program is the fall-back choice for analysis if regression is a failure (EC, 2004a–c).

Homoscedasticity of the residuals should be assessed using *Levene's test* as described in Appendix P, Section P.2.3, and by examining the graphs of the residuals against the actual and predicted values. Levene's test provides a definite indication of whether or not the data are homogeneous (as in Figure O.2A of Appendix O). If the data are heteroscedastic, then the graphs of the residuals should be examined. If there is a significant change in the variance and the graphs of the residuals produce a distinct fan or 'V' pattern (see Figure O.2B, Appendix O), then the analysis should be repeated using weighted regression. Before choosing the weighted regression, the standard error of the ICp should be compared to that derived from the unweighted regression. If the two standard errors differ by more than 10%, then the weighted regression is selected as the best choice [48]. However, if the difference is less than 10%, then the user should consult a statistician for the application of other models, or the data could be re-analyzed using linear interpolation (less desirable). This comparison between weighted and unweighted regression is completed for each of the selected models while proceeding through the process of selecting the final model and regression. Some non-divergent patterns might be indicative of an inappropriate or incorrect model (e.g., Figure O.2C, Appendix O), and the user is again urged to consult a statistician for guidance on other suitable models.

Choice of statistical software packages. The previous descriptions refer to use of a general-purpose statistical package (SYSTAT), but the future could bring "dedicated" software packages, designed for environmental toxicology. For example, the CETIS package contains an extremely wide choice of models for nonlinear regression. Jackman and Doe (2003) compared its estimates of endpoints with those from SYSTAT, for many models. In general, they found that the two software packages, and various models, produced similar estimates of EC20s, using a selection of real sublethal test results. However, they warned that results "often varied considerably" with different techniques, and some methods "gave results that were completely inappropriate".

Specifically, Jackman and Doe (2003) report that similar results were obtained with SYSTAT and CETIS for 13 sublethal data-sets with various organisms. In two other cases, the results from SYSTAT "seemed more reasonable", and in one case the reverse was true. They found that CETIS was more complicated and more difficult to learn than older toxicology packages. A "strong understanding of the statistical methodologies (or a very detailed guidance documentation) is required to make the correct statistical choices." They recommended good guidance for non-statisticians in selecting the proper nonlinear model from the wide choice available in the CETIS package. They also noted that if CETIS was used to estimate IC50s, it often did not provide reasonable estimates of confidence limits.

As more toxicology packages with nonlinear regression become available, it will be important to have guidance from a statistician in their use. It will also be desirable to compare estimates of endpoints from the new packages, with those obtained from general-purpose statistical packages using the standard method published by Environment Canada.

6.5.9 Newtox-Logstat—An Alternative Regression Program

Key Guidance

- *The Newtox-Logstat procedure obtains point estimates by regression. It has been successfully used in Canada for tests on inhibition of growth in duckweed.*
- *Newtox-Logstat offers investigators an alternative procedure for point estimates of sublethal quantitative data, at least for growth in plants. At present, it offers two models based on the Weibull and log-normal distributions. It does not model hormetic*

[48] The value of 10% is only a rule of thumb based upon experience. Objective tests for judging the improvement due to weighting are available, but beyond the scope of this document. Weighting should be used only when necessary, as the procedure can introduce additional complications to the modeling procedure. A statistician should be consulted when weighting is necessary.

effects, but might have broader capabilities in the future.

The toxicity analysis program *Newtox-Logstat* was developed at the Technical University of Denmark by Drs. K.O. Kusk and N. Nyholm, from a method described by Andersen (1994). (The main principles of a similar method were published by Andersen *et al.* (1998), although that publication has a different focus. An even earlier paper (Nyholm *et al.,* 1992) described the benefits of nonlinear regression for dealing with the statistical difficulties of quantitative data, and pointed towards the new method.) The program has been used in Canada, as described in the following text. The most convenient source of the program for Canadian investigators is from the Saskatchewan Research Council, by agreement of Drs. Kusk and Nyholm and Mary Moody[49].

The Newtox-Logstat procedure is suitable for sublethal quantitative results. It runs under an Excel spreadsheet, and each data-point is entered, not just the mean effects. It provides a choice of two nonlinear models based on the Weibull and normal log distributions. It estimates ICp and confidence limits.

The program was originally designed for data on growth rates. It has been successfully used in Canada by Moody (2003), for toxicity data on growth inhibition in duckweed (*Lemna* sp.; inhibition of increase in number of fronds and dry weight). Moody reports that the Weibull model provided the best fit to the data, visually. An example of the fit for inhibition of number of fronds is shown in Figure 17.

The Newtox-Logstat procedure offers Canadian investigators an alternative method of estimating endpoints by regression, certainly for tests involving algae and duckweed, and probably for other growth effects. At its present state of development, Newtox-Logstat does not have the capability of including hormesis in the model. The designers of the procedure suggested that hormetic effects should be arbitrarily set at zero inhibition for purposes of modelling. Moody (2003) found that hormesis did not generally interfere with analysis, but omitted such data when they did create a problem. [50]

6.5.10 General Linear Models

Key Guidance

- *General Linear Models and Generalized Linear Models represent broad categories of statistical models, including many familiar statistical techniques.*
- *GLIM is the more inclusive category, and covers a variety of distributions, including normal, exponential, probit, logistic, and Poisson. The approach could be applied to studies of quantal variables, or continuous ones such as weight.*
- *These concepts remain in the realms of statisticians at present, but have been advanced for use in toxicology. Benefits could be a single package of software to analyze diverse categories of results, transfer of knowledge and techniques among models, use of better mathematical methods instead of inexact techniques, and comparison of diverse models for their fit. However, non-statisticians would probably find the existing software packages difficult to use.*

---

[49] Ms. Mary Moody, Research Scientist, Environment and Minerals Branch, Saskatchewan Research Council, 125 -- 15 Innovation Boulevard, Saskatoon, Sask. S7N 2X8. (moody@src.sk.ca)

[50] From tests with frond inhibition, Moody (2003) took 23 endpoints derived by regression, and compared them to endpoints derived from smoothing and interpolation (ICPIN procedure). For the ratios of the endpoints (interpolation divided by regression) there was an overall similarity (average ratio 102%, median ratio 96%). However, there was great diversity in individual comparisons. The ratios ranged from 42 to 195%, with a standard deviation of 39%. Assuming that the endpoints from regression are more realistic, it represents an appreciable improvement in procedure. Comparison of dry weights of plants showed greater similarity between endpoints obtained by the two methods (standard deviation 20%), but in seven cases, interpolation failed to estimate the endpoint, or else failed to provide the confidence limits.

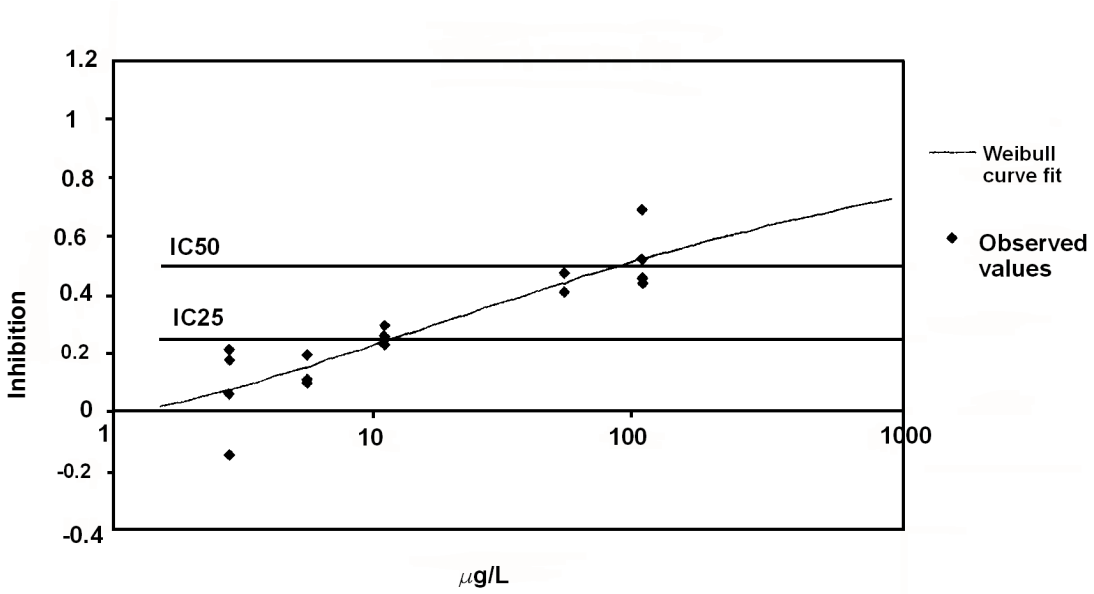


Figure 17 Effect of cadmium on inhibition of frond increase in *Lemna minor* (from Moody, 2003). The fitted curve is based on a Weibull model using the Newtox-Logstat procedures of Kusk and Nyholm developed from Anderson (1998).

The term *general linear model (GLM)* does not signify a specific technique, but a class or category of approaches, or a particular class of models. The models have a single dependent variable which is a function of an independent variable or variables. Thus simple linear regression falls into this class, but GLM should not be thought of as being limited to regressions. Also fitting the GLM class are models such as analysis of variance (ANOVA) and analysis of covariance (ANCOVA), which might not be thought of as "linear" models. Statisticians would point out that these procedures are "linear" because their parameters enter into the model in a linear fashion. Gad (1999) gives an example in which "the GLM procedure of SAS" is called up to do a conventional ANOVA for typical toxicological data (weight of kidneys as related to several doses).

Thus, investigators should expect many specific analytical techniques to be found under the broad umbrella of "general linear model". GLM is not at all a cut-and-dried computer package for simple application to a set of data. Most biologists or toxicologists would need direct participation of a statistician to apply these techniques to their work. GLMs are described by Searle (1971).

6.5.11 Generalized Linear Models

The term *generalized linear models* represents a larger category of mathematical models, that includes the GLMs of the previous section. The category has sometimes been called *generalized linear interactive models*, hence the acronym *GLIM (or sometimes GLiM)*.

The broad capabilities of GLIMs were initially useful as a teaching tool, but research and the advent of powerful computers have brought this category to the front and centre of statistical endeavours and developments. All GLIMs share the same mathematical approach, but the category could include a variety of specific statistical techniques. The techniques themselves might be of more direct interest in applied toxicology, than the more abstract mathematical concepts of GLIMs. An introduction to the topic is provided by Dobson (2002), and a more

detailed textbook was designed for statisticians and "numerate biologists" by McCullagh and Nelder (1994).

A variety of familiar mathematical distributions fit under the umbrella of GLIM, including normal, exponential, probit, logistic, and Poisson. These can be mathematically described so that an effect in any one of the distributions can be linked through a function to one or more independent variables. The effect might be quantal (counts, mortalities, proportions), or continuous variables such as weight. There is a common method for computing estimates of parameters. A researcher could use GLIMs to assess the dependence of an effect on a single independent variable such as concentration (by regression), or a more complicated structure of independent variables such as group treatment (ANOVA), or treatments and covariates (ANCOVA).

Enthusiastic support for GLIM in toxicological analyses is given in a series of papers by Bailer and Oris (1993; 1994; 1997) and associates (Bailer *et al.*, 2000a, b). They show that their general regression model can fit a variety of different effects, whether counted, dichotomous, or continuous. The regression can be used to estimate ICp, and circumvents the conceptual problems that exist in the computer program ICPIN (Bailer and Oris, 1997), being "superior to the [ICPIN] method in terms of bias, mean squared error, and coverage probability" (Bailer *et al.*, 2000b). Bailer and Oris (1994) comment that "computer software to fit the GLIM models ... is readily available (e.g., the GLIM macro in the NLIN procedure of SAS." In the earlier papers, confidence limits are not estimated, but Bailer and Oris (1997) list some options that could be developed with a defensible mathematical basis.

There are several benefits from use of GLIMs and their subcategory GLMs.

- A single package of statistical software can replace the array of programs needed to analyze non-normal and linear effects.
- The investigator's general knowledge can be transferred among the types of models under the GLIM umbrella (e.g., significance, goodness of fit, testing assumptions).
- Approaches that still incorporate inexact "short-cuts" and ad hoc techniques from pre-computer days can be abandoned in favour of better mathematical methods.
- It is straightforward to compare the fit to dose-effect data by various distributions (e.g., probit, logistic, Gompertz).

At the same time, there are limitations and drawbacks to GLIM approaches. Although a stand-alone software package for GLIM exists, biological investigators would probably find it difficult to use. Purchase of a larger package such as SAS would also provide GLIM capability, but personnel would have to know how to call up and master the appropriate techniques. Clearly, GLIMs are currently useful for toxicological research, but routine or regulatory toxicity tests will probably follow developed pathways, such as statistical guidance outlined in EC methods documents.

6.5.12 Reparameterization

This approach to analysis of toxicity data grows out of the desire to estimate endpoints and confidence limits in terms of a specific concentration (EC50, IC25), although the toxicity tests were set up with concentration as the *independent* variable. The degree of effect was actually the dependent variable, yet a fixed degree of effect is used to calculate the endpoint in terms of concentration. This "inversion" of the regression to select an endpoint entails some statistical complications, described in Section 9.4. One way of attempting to circumvent the complication is reparameterization to create a model containing the endpoint of interest. The approach was adopted by Stephenson *et al.* (2000), and has been modified as part of recent EC methods (Section 6.5.8).

Key Guidance

- *In environmental toxicity tests, the measured effect is the dependent variable. To calculate the endpoint, however, a fixed degree of effect is used as if it were the independent variable, in order to calculate the corresponding concentration of toxic material (the endpoint). This entails an "inversion" of the relationship.*

- *Reparameterization involves rewriting the statistical model describing the relationship to incorporate the endpoint (say the ICp) and its confidence limits as variables to be estimated by the model. This is done in linear/nonlinear regression techniques of Environment Canada (Appendix O).*
- *This procedure might result in poorer performance of the models, with the penalty of increased replication needed for satisfactory results.*
- *Other authors have published approaches to nonlinear modelling, and some examples are outlined briefly.*

Reparameterization starts with any standard statistical model, such as nonlinear regression. If the IC25 is to be estimated, the regression equation is "reparameterized" by rewriting it to include the IC25 as a parameter. This allows the IC25 and its confidence limits to be estimated directly, without the need for inverse regression techniques.

There are some drawbacks to the technique. In particular, statistical analysis might perform more poorly. For example, the hormetic logistic model (Section 6.5.8) was found to be very sensitive to the choice of optimization algorithm. It is therefore desirable to test more concentrations than might normally be used.

An early description of reparameterization was provided by Bruce and Versteeg (1992), who presented an excellent outline for using nonlinear regression on quantitative toxicity data. They tested the method on sublethal tests with the alga *Pseudokirchneriella subcapitata*, fathead minnows, and mysid shrimps. The resulting curves for measured effects at different log concentrations appeared to be smooth fits. The program then reparameterized the equation of the fitted line, to estimate the logarithmic ICp and its confidence limits, for any selected value of "p". Bruce and Verstag (1992) based their model "... on an S-shaped curve derived from the cumulative normal distribution", and they provide the code for carrying out the analysis with SAS. Another example is provided by Andersen *et al.* (1998).

This procedure has been incorporated into the models offered for the new EC test methods (EC, 2004a–c), described in Section 6.5.8 and Appendix O.

6.5.13 Other Examples of Regression Trials

Regression methods used by some other authors are outlined here. The procedures look promising, but require knowledgeable application.

A family of nonlinear models, similar to the ones discussed previously, has been described by Slob (2002). Analytical procedures are carried out by an "easy-to-use" software package called PROAST, which is available within Slob's institute in the Netherlands. One of the good features of the regressions is the determination of the *Critical Effect Dose* (CED), which is related to a negligible or acceptable degree of effect on the test organisms.

Generalized nonlinear regression was recommended for estimating ICp and its confidence limits by Andersen *et al.* (1998). A plot of data was used to choose a particular regression function. For their analysis, the authors amalgamated some standard numerical routines, including coding in FORTRAN 90. Their method used "... the empirical nonhomogeneous variance and covariance in the estimation of the dose-response curve". A version operated in Windows 95 format.

Scholze *et al.* (2001) used 10 different sigmoidal regression functions, the more familiar being probit, logit, Weibull, generalized logit, and three options of Box-Cox functions. All functions were fitted to a given set of data, and the best was selected by two stages of testing (residuals, then goodness of fit). Bootstrap estimates provided confidence limits. The method was tested by a remarkable prediction of total toxicity of a mixture of 14 substances with differing modes of action. The predicted effect on inhibition of bacterial luminescence was 36%, almost identical to the actual observation of 39%.

Moore and Caux (1997) used five "generic" models on quantal and quantitative data. Best fits were usually obtained by a *three-parameter logistic equation* with a steep slope parameter. They also tried three logistic models, a two-parameter probit model, and a two-parameter Weibull model. Higher order polynomials, which had little biological plausibility, were excluded. Their package ran in a

spreadsheet format, used logarithm of concentration, and gave a maximum-likelihood fit with each model (Caux and Moore, 1997). Output included goodness of fit, graphs with observed data and fitted curve, and ECps or ICps from low to high values of "p". From 198 sets of sublethal data, they selected 65 sets which had reasonable monotonic dose-effect relationships and at least one partial effect. They analyzed the 65 sets using their method, and claimed an adequate fit in about 40 cases.

Baird *et al.* (1995) claimed that only two nonlinear parametric models were needed to deal with diverse toxicity results. Using a *logistic dose-response* and a *power model*, they fit 77–100% of quantitative sublethal tests with minnows, sea urchin, abalone, and giant kelp. The power model had the form $y = bx^c$ and would fit straight lines and upwardly concave or convex distributions. However, their validation was uncertain, because their hypothetical data had very unrealistic arithmetic ranges, and logarithmic data were analyzed using arithmetic concentrations. Data for giant kelp were fitted and graphed as a curve, but would apparently have been a straight line if a proper logarithmic scale of concentration had been used.

6.6 Thresholds from Regression

Key Guidance

- *There is some international movement to develop methods which would estimate the "true" or absolute no-effect concentration for a population of organisms. That would be a theoretical value and would have to be estimated using regression techniques.*
- *Workers in the Netherlands have developed such models to estimate the No-effect Concentration or the Critical Effect Dose.*
- *"Hockey-stick" models estimate such a threshold of effect. The long "handle" is the normal concentration-effect regression, while the "blade" represents the background of normal effects. The intersection is taken to represent a threshold.*

There is a current thrust in Europe, and internationally, to develop toxicity models that estimate the "true" *no-effect concentration* (NEC) (OECD, 2004). That estimate would be an absolute no-effect level, which is a parameter of the population, not the sample that is tested (Anon., 1994). The goal would be approached by regression techniques, not hypothesis testing which estimates the *observed* no-effect concentration (NOEC) in a sample, rather than the true one. Examples of these developments in Europe are given in Section 6.6.2.

6.6.1 Thresholds with the Hockey-stick Model

Threshold modelling for quantitative sublethal toxicity can be done using a "hockey-stick" model. This is a linear model for regression since two straight lines are fitted to data from the test. The longer "handle" of the model applies to the usual dose-effect relationship, and the "blade" of the model would be a line parallel to the concentration axis (Figure 18). Zajdlik (1996) works through the mathematical steps of fitting such a model, and comments that it is not difficult.

Zajdlik's description lends considerable appeal to the approach, which estimates an apparently objective and meaningful threshold of effect as the "join point" of the two lines. He points out some potential drawbacks, such as the general problem of fitting a curved dose-effect relationship. Sometimes a particular toxicant might not demonstrate a threshold of effect (lower concentrations would simply take longer to act). Zajdlik (1996) indicates that it might be more expensive to run the experiment for this kind of analysis, but that would be balanced by the advantages of an objective estimate of an environmentally "safe" concentration.

The method discussed by Zajdlik (1996) has been used by other authors. One excellent example is for incidence of liver lesions in benthic fish, as related to PAHs in the sediments (Horness *et al.*, 1998). Graphical representation (Figure 18) shows background incidence of lesions distributed horizontally along one segment of the hockey stick, across a range of low logarithmic concentrations of PAH. Then there is an abrupt change as the second segment of the regression shows a linear increase of lesions with higher log concentrations. The fits appear to be reasonable, although there are rather large confidence intervals for the relationships in panels A and D. Confidence intervals are not indicated for panel B, but might also be large. Still, the estimate of an apparent threshold for toxic effects

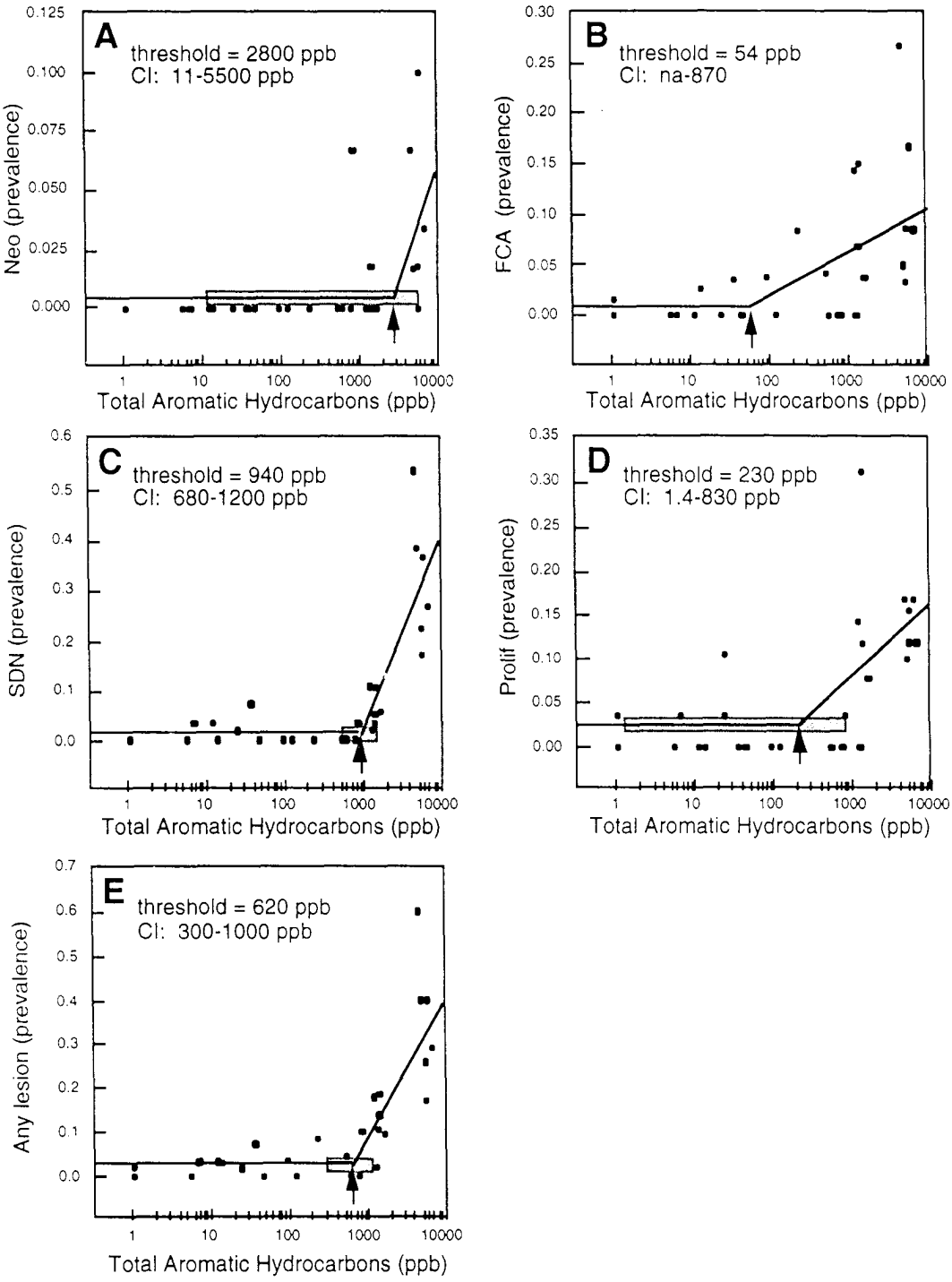


Figure 18 **Examples of hockey-stick regression** (from Horness *et al.*, 1998)**.** The panels represent data for types of hepatic lesions in English sole, collected at Pacific coast locations. Vertical scales represent frequency of occurrence among fish. Horizontal axes are measurements of total aromatic hydrocarbons in dried bottom sediment from the same locations. Threshold concentrations are indicated by arrows, and the shaded bars represent confidence intervals.

at the junction of two segments appears to be a very useful piece of information.

For their analysis, Horness *et al.* (1998) treated the two segments (lines) as a single discontinuous function, defined by a single regression. The PAH concentrations were transformed to logarithms before analysis, although it could be done within the calculations. Horness *et al.* point out that iterative numerical techniques for estimating nonlinear regression parameters are increasingly available in standard commercial packages, and they used the SAS statistical package JMP®.

The potential usefulness of "hockey-stick endpoints" is forcefully demonstrated by Beyers *et al.* (1994), who estimated thresholds of toxicity that were 2- to 4-fold lower than the NOEC derived by hypothesis testing. They studied toxicity of pesticides to fish, and their fits to the hockey-stick model appear to be satisfactory. They also used the statistical software developed by SAS.

6.6.2 No-effect by Regression

Nonlinear regression should be amenable to estimating thresholds of toxic effect, and workers in the Netherlands have taken that approach. Slob (2002) demonstrated the use of a family of nonlinear regressions to determine the *Critical Effect Dose* (CED), related to a negligible degree of effect on the test organisms (see Section 6.5.13).

Similarly, Kooijman and Bedaux (1996) offer a description and software for estimating the sublethal endpoint designated as the No-Effect Concentration (NEC). Their program is designed primarily for analysis of the sublethal tests methods published by the OECD, on growth of fish, reproduction of *Daphnia*, and algal growth. The authors indicate that the program can also yield analyses of quantal data on mortality (LC50s), effective concentrations (EC50s), and effective times (ET50s), all with confidence limits. These claims have not been validated for the present Environment Canada document.

Kooijman and Bedaux (1996) supplied the computer program on a disk *(DEBtox*, signifying Dynamic Energy Budget), as part of a book. The program published in 1996 operated under Windows 3.1 or 95). More recent versions operate under Windows and Unix, and are available on the Internet at: www.bio.vu.nl/thb/deb/deblab/. The program features were described in detail recently in an OECD guidance document (OECD, 2004). The program appears to be well designed, clear, and easily used. Example data run easily under the program; it produces endpoints and supporting information, but does not give indications of just what models and procedures were used to obtain the answers.[51] The program offers graphs which can be printed out if desired. Unfortunately the concentrations were plotted on an arithmetic scale which provides the viewer with a distorted impression of asymptotes, apparent thresholds, and general shape of the curves.

The NEC approach is also embodied in a mathematical function for inhibition of population growth in algal tests (Kooijman *et al.*, 1996). The equation is reported to perform well, equalling the effectiveness of logistic, log-normal, or Weibull analyses (N. Nyholm, 2001, personal communication, Tech. University of Denmark, Lyngby, Denmark).

The advantages of the NEC approach are obvious. It uses suitable statistical procedures, namely fitting a regression. It satisfies the demand for an endpoint that represents the threshold of effect, ostensibly no effect.

[51] An attempt at entering new data failed. The Canadian operator was able to enter numbers in some positions of the initial data-table, but was unable to discern which parts of the table were to receive data on concentration, time, number of test organisms, and effect. He was unable to find guidance on the issues.

Hypothesis Testing to Determine NOEC/LOEC

7.1 General Suitability for Environmental Testing

Key Guidance

- *Hypothesis testing determines statistically significant differences between results for the control and results at each test concentration.*
- *This is one approach for single-concentration tests such as those used in monitoring.*
- *In a multi-concentration test, hypothesis testing identifies the no-observed-effect concentration (NOEC) and lowest-observed-effect concentration (LOEC).*
- *Estimate of NOEC/LOEC is an option in some sublethal test methods published by Environment Canada. However, it does not represent a good toxicological endpoint in multi-concentration tests, for several reasons including the following ones.*
  - *The endpoints are defined statistically rather than biologically; higher variability within the test leads to higher values of NOEC/LOEC.*
  - *The NOEC does not necessarily represent a safe level in the environment although it conveys that impression.*
  - *The endpoints can only be concentrations that were actually tested and are therefore open to manipulation by chance or design.*
  - *The calculations produce a pair of concentrations, rather than one endpoint.*
  - *No confidence limits can be calculated.*
- *The geometric mean of NOEC and LOEC can be used to provide a single endpoint, and should be called the threshold-observed-effect concentration (TOEC). It has the same shortcomings as NOEC and LOEC.*

7.1.1 Single-concentration Tests

Hypothesis testing is standard procedure for toxicity tests which have replicates of one concentration and a control (e.g., samples of sediment from one location). This is a suitable statistical approach, and there is no other. Available techniques are described in Section 3.

Comparison of one treatment with a control can be made with a *t-test.* Multiple t-tests must not be repeated in a set of samples, in lieu of a multiple-comparison test. Special modifications of the t-test are available (Appendix P.4.4).

The remainder of this section deals with tests which have at least two test concentrations or sets of samples.

7.1.2 Multi-concentration Tests

Point estimates such as IC25 are recommended in this guidance document as the primary endpoint, and hypothesis testing is assigned a secondary position. However, several EC methods documents allow hypothesis testing to be used if desired. Accordingly, the procedures are outlined here, since they might be relevant to a particular situation, and also, to allow evaluation of existing work which used this statistical method.

The variables estimated in hypothesis testing would be the *no-observed-effect concentration (NOEC)* and *lowest-observed-effect concentration (LOEC)*. The usual method for determining NOEC and LOEC is to statistically compare the control effect to the effects at individual test concentrations (see Sections 7.4 and 7.5). Hypothesis testing is frequently used, in part because there are well-established methods. The ANOVA and nonparametric methods are readily available, relatively easy to use, and robust in the face of irregular data. However, there is a growing literature which points out many deficiencies of the hypothesis testing approach (Suter *et al.*, 1987; Miller *et al.*, 1993; Pack, 1993; Noppert *et al.*, 1994; Chapman, 1996; Chapman *et al.*, 1996b; Pack, 1996; Suter, 1996; Moore and Caux, 1997; Bailer and Oris, 1999; Andersen *et al.*, 2000; Crane and Newman,

2000; Crane and Godolphin, 2000). Canadian environmental toxicologists and statisticians are among those with misgivings (Miller *et al.*, 1993). Several limitations are listed in the following text.

- NOEC and LOEC can only have values equal to concentrations that were actually tested. Since that was determined by the investigator, the endpoints could be susceptible to chance influences, whim, or manipulation.

- NOEC and LOEC are particularly sensitive to variability within the test, since they depend on determining statistically significant difference from the control effect. A careful test with a precise result would result in a lower NOEC, while a test with great variability would result in a higher NOEC. Thus, the endpoints NOEC/LOEC do not relate to any particular point on the dose-effect curve.

- The values of NOEC and LOEC depend to some extent on the method of statistical analysis that is used.

- The statistical power of the ANOVA and multiple-comparison test is often low because of relatively few replicates. Fewer replicates result in a higher NOEC, so there could be an incentive to either reduce or increase the replication, depending on orientation of the program.

  [This situation could be remedied by increased consideration of the power characteristics in design of the test (Section 7.23). There could be a requirement imposed, to demonstrate that a test had adequate power to detect, say, a 25% effect.]

- No confidence limits can be calculated for the endpoint; therefore, different NOECs cannot be compared statistically.

- The name of the NOEC has some popular appeal, and might be mistaken for a "safe" concentration by non-toxicologists, even though it can be associated with appreciable effects.

- Estimating an NOEC tends to be somewhat in conflict with a basic rule of the scientific method, because there is an attempt to "prove" a null hypothesis of no effect.

In fairness, it should be pointed out that these limitations are not unique to hypothesis testing. Most of them have parallel weaknesses in other approaches used to analyze test results. For example, the conventional confidence intervals for IC25 and EC25 depend on the assumption that the model used to produce them is correct. That assumption is rarely even acknowledged, let alone tested.

The importance of precision of results and choice of statistical method is illustrated in an example by Crane and Godolphin (2000). They present some hypothetical data for lethal testing of the same effluent by laboratory no. 1 which obtained precise results, and laboratory 2 which had variable results. Analysis by Dunnett's test (currently the most popular multiple-comparison test) yielded NOECs of 2.2% for laboratory 1, and 22% for laboratory 2. Choice of other statistical tests produced great variation in estimated NOEC, from 1.0 to 10% for results of laboratory 1, and from 2.2 to 46% for laboratory 2 (see Section 7.5.1).

Other specific examples can be given. Suter *et al.* (1987) demonstrated that estimation of NOEC/LOEC did not provide satisfactory endpoints. When sublethal studies on fish were analyzed with nonlinear regression, a comparison with results of hypothesis testing showed that the geometric means of NOEC and LOEC *(TOECs* see following text) were associated with effects of 12% on hatching, 19% on larval survival, 20% on parental survival and larval weight, 35% on weight per egg, and 42% on fecundity. Those are relatively strong effects, certainly demonstrating that TOECs can be far from a true threshold of effect. A similar analysis of 14 sublethal test results showed that the NOEC (not the TOEC), was associated with sublethal effects which ranged from 3 to 38%, with a geometric average effect of 14% (Crane and Newman, 2000).

Concerning the appeal of the name NOEC, the minutes of the meeting in Quebec City noted hypothesis testing "has an appeal from regulatory and management perspectives" because it sounds as if it answers the question about whether a given concentration in the environment is, or is not, toxic (Miller *et al.*, 1993). Environmental toxicologists will be aware that any answer to that question by determining NOEC/LOEC could be erroneous because of the problems itemized previously.

7.1.3 Expressing Results as a Threshold

A geometric mean of the NOEC and LOEC is often calculated for the convenience of having one number rather than two. A recommended term for this geometric average is the *TOEC* signifying *threshold-observed-effect concentration*. The use of "threshold" is in the dictionary sense of "point at which an effect begins to be produced". Such a value may be used and reported, recognizing that it represents an arbitrary estimate of an effect-threshold that could lie anywhere in the range between the LOEC and NOEC, and is subject to all the uncertainties of those values (Section 7.1.2).

The term *Maximum Acceptable Toxicant Concentration (MATC)* was used in the past, mostly in the USA, as an empirical endpoint for sublethal life-cycle tests. Various authors used the term confusingly, to signify (a) the geometric mean of NOEC and LOEC, (b) the NOEC, (c) a value that could not be determined, between NOEC and LOEC, or (d) the range from NOEC to LOEC. Recent literature tends to abandon the linguistically tormented term MATC giving preference to the use of NOEC and LOEC; TOEC is recommended here. Point estimates are more appropriate for determining thresholds (Section 4).

7.2 Design Features in Hypothesis Testing

Key Guidance

- *Major errors in analysis of variance could result from using non-replicate measurements as if they were replicates.*
- *"Type I error" means a "false positive", i.e., concluding that a difference existed between treatments when there was actually no real difference. "Type II error" means accepting a null hypothesis of no difference, although a real difference existed.*
- *Most investigators set the level of significance (α) as probability (p) = 0.05. As a result, 5% of toxicity tests can be expected to show a difference by chance alone, leading to a one-in-twenty possibility of a Type I error.*
- *The probability of making a Type II error is β (Beta). It is inversely related to α, so if lower p-values are selected, there is greater likelihood of a Type II error. The power of a test is (1 - β), and is the discriminating ability of a test. Most investigators do not design tests in terms of β or power, although it would be desirable.*
- *If applied to real environmental situations, a Type I error in toxicity testing would result in stricter discharge limits or additional waste treatment, which might not be necessary. A Type II error would increase the chance of damage to the environment.*
- *The Minimum Significant Difference (MSD) should be reported as an alternative way of describing the power of a toxicity test. The MSD is the smallest percentage difference between results for the control and a treatment, that would be statistically different within the design of the toxicity test.*
- *"Bioequivalence" is a reverse application of MSD; before starting the toxicity test, a degree of acceptable difference between treatment and control is set for a "pass" in the test.*
- *Hypothesis testing should not normally be used on untransformed quantal data. It can be used, however, if replicates have quantal observations on ≥100 individuals; the numbers are great enough that they approach a quantitative distribution. This approach is currently seeing some use, although superior methods might well be required in the future.*

7.2.1 Replicates and Experimental Units

In hypothesis testing, it is very important to identify experimental units and the true replicates (explanation in Section 2.5 and warning note here). A sloppy designation of *replicates* could lead to extremely erroneous analysis and conclusions. In particular, organisms in a single test container would not represent replicates, but *sampling units*.

7.2.2 Errors of Types I and II

In hypothesis testing, it is particularly easy to arrive at mistaken conclusions in a manner that is either overly optimistic, or else too conservative. The topic is closely related to the *power* of the statistical test used in testing a hypothesis (see Section 7.2.3). The topic

Table 4 Types of errors in hypothesis testing, with associated probabilities (after USEPA and USACE, 1994).

Conclusion drawn from hypothesis testing	Actual (true) status of the populations	
	No difference (H_0 is true)	Difference (H_0 is false)
Accept null hypothesis (H_0 concluded to be true)	Correct (Probability = 1 - **α**)	Type II error (Probability = **ß**) a "false negative"
Reject null hypothesis (H_0 concluded to be false)	Type I error (Probability = Significance level = **α**) a "false positive"	Correct (Probability = Power = 1 - **ß**)

is also related to questions of statistical versus biological significance (Section 9.3).

A *Type I error* (a "false positive") occurs when a null hypothesis that is actually true is rejected (i.e., a difference is concluded when there is none). A Type II error (a "false negative") occurs when the null hypothesis of no difference is accepted (= not rejected), even though it is actually false and a difference exists. Table 4 shows the relations between test conclusions and the (unknown) true situation.

Most investigators partially control these mistakes by setting the level of significance ("**α**") for tolerating false positive results (Type I error). Almost always, **α** is set so that the probability *(p)* = 0.05. If that is done, one test out of 20 (5% or 0.05) can be expected to show an apparently significant difference by chance alone, i.e., the items being compared are strongly divergent, but not really different. Accordingly, there is a one-in-twenty chance of concluding a "false positive" or Type I error. If a higher value of **α** were selected (say, 0.1), the chance of concluding a false positive would increase (in this example, a chance difference would be expected in one of ten trials). At a low value of **α** = 0.01, only one in a hundred trials would be expected to show a Type I error (but see the following text for the penalty).

The probability of making a Type II error is called ***ß*** *(Beta)*, the probability of accepting the null hypothesis when it is actually false (Table 4). The value of **ß** is seldom set deliberately by the investigator before doing the test (see following text), but is determined in large part by the initial choice of α. There is an inverse relation between **α** and **ß**, and as the significance level is lowered, (smaller selected value of **α** and less possibility of a Type I error), Beta becomes larger, so there is greater likelihood of a Type II error. As another contributing factor, the more powerful the design of the test (e.g., more replicates, Section 7.2.3), the less likelihood there is of making a Type II error.

Statisticians usually talk about Beta in terms of the *power* of a test, which is (1 - **ß**), and can be defined as

(a) the "discriminating ability" of a test,
(b) the probability of correctly concluding that there is a difference, or more properly,
(c) "the probability of rejecting the null hypothesis when it is in fact false and should be rejected".

When applying the results of toxicity tests to the real world, there are very different implications for making Type I and Type II errors. Falsely concluding that there is a toxic effect (Type I), if applied to an industrial discharge or to the setting of water quality limits, could lead to tightening up restrictions, or applying further waste treatment. The consequences would be a wider margin of safety for the natural world, and increased cost in human activity [52]. On the

[52] The importance of the selected level of statistical significance and power was demonstrated in a surprising result from Moore *et al.* (2000). They submitted samples of non-toxic, laboratory-manufactured water to testing

other hand, failing to detect a difference which was actually real (Type II error) would create an unwarranted impression of safety for the tested material, possibly resulting in a receiving environment that damaged organisms. From the ecological point of view, Type II errors are more serious. Accordingly, the significance level (α) should not be set at unduly strict levels. Choosing a significance level of 0.01 instead of 0.05 might seem to be setting a high standard, but it would also decrease the power of the test, increase the likelihood of a Type II error, and increase the possibility of harmful consequences for the environment.

7.2.3 Power of a Toxicity Test

Section 7.2.2 introduced the concept of the *power* of a statistical test in hypothesis testing. The power is influenced by several factors:

- the significance level (α) chosen by the investigator;
- variability of replicates;
- effect size *(ES*, the magnitude of the true effect that is being tested); and
- *n*, the number of samples or replicates used in a test.

Power analysis can be used *a priori* to determine the magnitude of the Type II error and the probability of false negative results (USEPA and USACE, 1994). Three of the four listed items can be selected by the investigator and incorporated into the design of the test. The fourth item, variability, is difficult to predict, but can be estimated from past experiments, or trial tests. Accordingly, it could be time-consuming or tedious to design power into a test, so it is not often done. Designing for suitable power might indicate the need for a large test, economically and logistically unattractive. In that case, investigators should at least recognize the limitations and the possibility of an incorrect conclusion.

No standard value has been developed for the power of a test, or for its basis, which is the Type II error rate, **ß**. A value of 10% for **ß** (power = 90%) has been adopted for monitoring of effects at metal mines (EC, 2002b), and might be considered a suitable goal. However, even at this power, any conclusion about lack of toxic effect could be shaky. At a power of 90%, one test in ten might fail to show an effect because of chance, perhaps related to small sample size or variability among the organisms. It is wise to temper conclusions of "no effect" with the qualification "for this design and power of test". For a low-power test, it could be more realistic to report an inconclusive result rather than no effect.

Toxicologists have been urged to report both α and statistical power ($1 - \beta$), as indications of the possibilities of drawing false conclusions in either direction. Most people have difficulty with power, and do not report it. Indeed, it is a fairly complex item, with different specific formulae for various statistical tests. In view of the complexity, an alternative approach using Minimum Significant Difference is outlined in Section 7.2.4. Investigators wishing to report power of toxicity experiments should consult USEPA and USACE (1994).

7.2.4 Minimum Significant Difference

The *Minimum Significant Difference* (MSD) is a particular case of the power in a given test, and can be regarded as "an index of power". Since MSD is a feature of the software for many multiple-comparison tests (Section 7.5), reporting it is a partial remedy for the difficulty in communicating the power of a toxicity test.

The exact meaning of *Minimum Significant Difference* depends on the statistical test being considered. In general, MSD is the magnitude of the difference that would have to exist in average measurements (weights, for example), between the control and a test concentration, in order to conclude that there was a significant effect at that concentration. Clearly, MSD becomes larger with increased variation within concentrations.

The MSD is often stated as a percentage. For example, an MSD of 12% would mean that a difference of 12% between the measurements for a test concentration and those for the control, would be the minimum difference that could be detected in the toxicity test. (In other words, if a 12% difference were found, it would be considered statistically

---

laboratories, as supposed samples of wastewater. Six of 14 laboratories reported that the water was toxic. Moore *et al.* (2000) could not find any plausible reason for this high level of Type 1 error, and suggest remedies including additional criteria for acceptance of toxicity tests.

significant, for the manner in which the test was conducted).

If NOEC/LOEC are reported as endpoints, it is beneficial to report the MSD. The user of the results will gain some sense of the variability in a given test, and how closely the result should be interpreted. Environment Canada requires the MSD to be stated in reports under their programs, for the statistical tests in which it is available. It is strongly recommended that the MSD should also be stated in reports under other jurisdictions (Miller *et al.*, 1993).

The MSD or its equivalent is provided by all parametric multiple-comparison tests, such as Williams' and Dunnett's tests (Section 7.5). Unfortunately, in today's common practice, there is no useful analogue provided by nonparametric tests (such as Steel's Many-One or Wilcoxon Rank Sum tests.)

Acceptable values for MSD. As yet, Environment Canada has no guidelines for deciding an acceptable MSD.

Attendees at the meeting of the Statistical Advisory Group considered adopting a cutoff point, such as invalidating a test with MSD greater than 50%, but no decision was reached (Miller *et al.*, 1993).

Washington State has adopted an MSD of 40% in sublethal tests, for regulatory purposes (WSDOE, 1998). The USEPA (2000b) has offered some recommended maxima for accepting results in certain toxicity tests (Table 5). The values were derived by inspecting a national data base for tests with reference toxicants, for 23 test procedures at 75 laboratories over 10 years. The maxima would apply no matter what probability value (α) was selected.

From Table 5, it appears that the normal value of MSD is very much an individual characteristic of various toxicity tests. Apparently, it is not appropriate to have one value of MSD for all organisms and procedures. In dual-effect tests, the recommended values apply only to the sublethal effect.

The same conclusion of the need for different MSDs for different tests was reached in an objective study by Wang *et al.* (2000). From trials with suitable sets of data, Wang and co-workers concluded that MSD limits could be set in a scientifically sound manner. The limit would be chosen from a fairly complex equation, which they provided, and depended on several other variables including power of the statistical test and the desired detectable difference from the control. No single "cook-book" value could be given.

7.2.5 Bioequivalence

Bioequivalence is the name which has been given to a testing approach related to MSD. This tool for hypothesis testing has the effect of turning the general approach and use of a null hypothesis upside down. First, a degree of acceptable difference is set, between the performance of the control and the test concentrations. The null hypothesis is that test results are not "further out" than the acceptable difference. The hypothesis is tested by statistical treatment.

Shukla *et al.* (2000) show advantages of using bioequivalence. Many toxicity tests with an appreciable effect which had received a "pass" under the conventional approach, because of high variability within the test, showed a (deserved) failure under the bioequivalence approach. Many tests with only a slight toxic effect, which had shown a "failure" of the test material under the conventional approach, because of slight variability within the test, showed a (deserved) "pass" under the bioequivalence approach. Statistical background for the bioequivalence approach is provided in Wellek (2002).

Use of the bioequivalence approach requires agreement on what is a biological meaningful effect, which is not decided for most EC tests (see previous text). However, there have been some initial approaches to limits of acceptable effect in Canadian tests. Regulatory control of industrial effluent usually requires that a test of acute lethality must show an effect of less than 50%. This does not imply that killing almost half of the test organisms is acceptable. The endpoint was adopted as one that could be estimated with reasonable accuracy and dependability. Beyond that, the philosophy was that reasonable control of toxicity in the discharged waste itself, would achieve satisfactory conditions after dilution in the receiving environment.

The Canadian Disposal at Sea Program has somewhat more restrictive requirements for two toxicity tests.

Table 5 Minimum significant differences recommended by the USEPA for certain sublethal effects in selected toxicity tests (after USEPA, 2000b).

Test method published by the USEPA	Effect measured	Maximum MSD
Ceriodaphnia, reproduction and survival	Reproduction	37%
Fathead minnow, larval survival and growth	Growth	35%
Inland silverside, larval survival and growth	Growth	35%
Mysid, survival, growth, fecundity	Growth	32%
Sheepshead minnow, larval survival and growth	Growth	23%
Pseudokirchneriella subcapitata, growth/multiplication	Growth/multiplication	23%

In the echinoid test (EC, 1992f), a sediment fails if fertilization success is 25% less than the success rate in control water. In the marine amphipod test (EC, 1992d; 1998b), a sediment fails if survival is 20% less than the survival rate in a reference sediment, or 30% less than in a control sediment (Porebski and Osborne, 1998; Zajdlik *et al.*, 2000). There is also a requirement that the difference should be statistically significant. In other words, the apparently deleterious effect should not be the result of chance. The criteria for a valid toxicity test must also be satisfied. The scientists of the Disposal at Sea Program intended these limits to be reasonably representative of an ecologically significant difference from natural variability in populations. They were strongly aware of the restricted knowledge for setting such limits, but clearly the limits were required for regulatory programs. Validation of the choice is a topic of investigation (Zajdlik *et al.*, 2000).

7.2.6 Using the Techniques on Quantal Data

It is possible to use hypothesis testing, normally a quantitative technique, for evaluating quantal effects, but usually, that should not be done. One exception, however, would be if the data were suitably transformed, as outlined in Sections 2.9.2 and 2.9.3.

Another exception would be if each replicate had $\geq$100 observations; quantitative analysis could be used directly, as discussed in Section 6.1.1. An example is the echinoid fertilization test (EC, 1992f), with quantal data for fertilization of 100–200 eggs per container, it is satisfactory to treat the data as if they represented a continuous distribution.

If replicates have low numbers of individuals, say less than 100, results must be analyzed as quantal data.

7.3 Preparation for Testing by ANOVA

Hypothesis testing is a well-recognized procedure, with a general approach that is commonly used for research in pharmacology and human health. The approach has a set of statistical techniques which can be used when the data are quantitative, i.e., variable in magnitude among individuals, such as size, weight, or number of tumours [53].

[53] Some tests with quantal data can be analyzed by hypothesis testing if the numbers of observations are large (Section 7.2.6).

Key Guidance

- *For hypothesis testing of quantitative results, the effects of exposure to different treatments are examined for statistical differences. There must be replicates. Often, the different treatments would be a series of concentrations and a control (assumed in the following text).*
- *The Shapiro-Wilk's test is used to assess normal distribution of the data, and O'Brien's test (or Levene's or Bartlett's test) judges homogeneity of variances in the various treatments. With favourable results, the investigator can proceed to parametric analyses. Plotting a graph could help to assess normality.*
- *If data do not conform to normality and homogeneity of variance, they might be made to do so by transformation of the data. Analysis could then proceed by conventional parametric methods.*
- *If transformed data still do not conform to normality and homogeneity of variance, nonparametric methods of analyses must be used. Parametric analyses would not be valid, but might also proceed, to compare the estimated sensitivities. The parametric methods are relatively robust for small deviations from normality and homogeneity of variance, and for such mild deviation, results from parametric analysis might be reported in addition to the required non-parametric analysis.*

The fundamental approach is to adopt a *null hypothesis* that the effects shown by organisms in the test concentrations will be no different from those seen in control organisms. The toxicity test is then run, and the degrees of effect are measured in replicated groups of test organisms at one or more concentrations, and in control organisms. When parametric methods are used, the statistical comparison of the degrees of effect reveals whether the differences between differently treated groups *(variation among*) are statistically greater than the overall *variation within* each treatment. In nonparametric methods the comparison is based on the relative rankings among treatments. If there is no effect of treatment, the average ranking should be the same for the various treatments. If no difference is detected between any test concentration and the control, compared to the general "noise" among replicates, then the investigator accepts the null hypothesis, i.e., no effect of the test condition(s). If there is/are significant difference(s) between treatment(s) and control, the null hypothesis is rejected, and automatically the *alternative hypothesis* is accepted, that there is a real effect of the test material, i.e., toxicity.

The general statistical procedures in environmental toxicology were well developed in the 1980s and 1990s (citations in Appendix P), and statistical background is provided by Wellek (2002). The general flow of hypothesis testing is indicated in Figure 19. Usually, if a toxicity test is suitably designed and elicits consistent effects among the test organisms, it will follow a track vertically downwards in the centre of Figure 19, while tests that have some irregularity or problem will divert towards the right side. The most-recommended multiple-comparison tests are indicated at the bottom of Figure 19, along with substitutes if the first choice is not available. Some others are mentioned in the text.

An investigator should plot the results from a test, even though hypothesis testing does not fit a line to the data. Examination of the plot allows one to assess whether NOEC and LOEC are reasonable, and to see any anomalies in the data (see examples in Section 10.4).

Although test concentrations should have been selected in a geometric series (Section 2.2), under usual circumstances the scale of concentration is not a factor in the statistical analysis, which deals with effects. The concentrations serve only as a label for the groups.

7.3.1 Tests of Normality and Homogeneity of Variance

Analysis of variance (ANOVA) is at the core of hypothesis testing in parametric analysis. It is based on assumptions that data are normally distributed, and that variance is similar in different groups/treatments. The same assumptions apply to the parametric multiple-comparison tests that follow

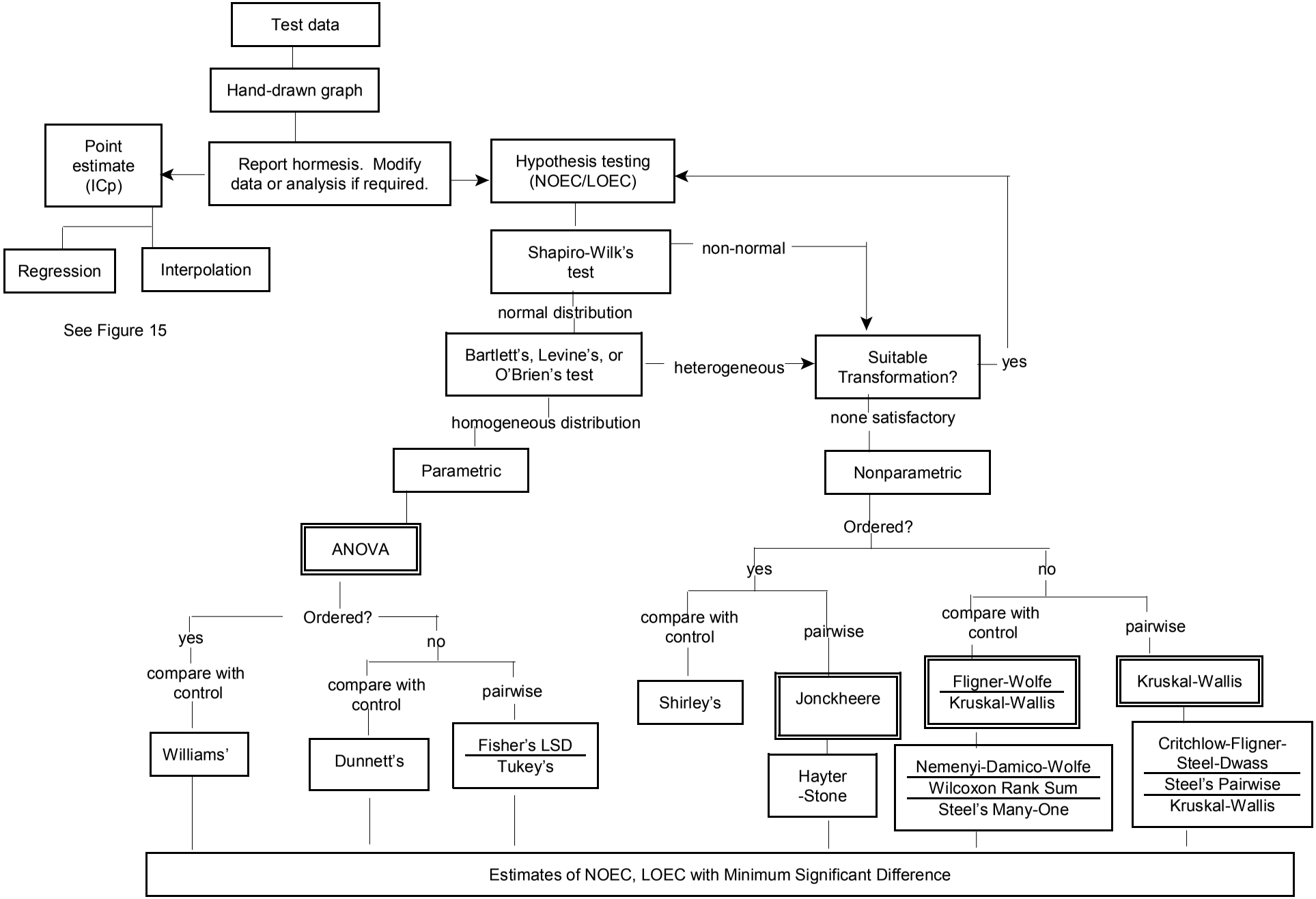


Figure 19 **Sequence of statistical analyses for testing hypotheses in toxicity tests.** A box with double outline indicates a test of a null hypothesis. Only if that hypothesis is rejected, does analysis proceed to the multiple-comparison test.

the ANOVA. The investigator must test whether those assumptions are met, before using ANOVA. The tests are listed in this section and described more fully in Appendix P, Section P.2. There must be at least two replicates for all of these statistical tests and more are desirable; the deficiencies of the multi-comparison tests become more serious if the number of replicates is small.

If one or other of the qualifying tests is failed, the data must by analyzed by the alternative nonparametric methods (Section 7.5.2). For mild non-conformity, there could be some benefit in carrying out and reporting both parametric and nonparametric analyses (Section 7.3.2).

Normality. The *Shapiro-Wilk's Test* is recommended for testing normality, rather than the *Kolmogoroff-Smirnov Test* available in some computer programs. Shapiro-Wilk's is described in Appendix P, Section P.2.1, and an example is given. The analysis is based on residuals, with a minimum sample size of three. Standard toxicological computer programs handle the complicated calculations. Final comparison uses a critical value (W), provided in tables (Shapiro and Wilk, 1965; D'Agostino, 1986), and investigators can assess the degree of non-conformity.

In addition, plotting the data for each replicate or concentration could be instructive (see Figure P.1, Appendix P). The graph could suggest the apparent cause of non-normality or non-homogeneity.

Homogeneity of variance. The *test of Levene* (1960) is recommended here, but unfortunately it is not included in the software packages designed for environmental toxicology. Levene's test avoids a problem shown by Bartlett's test, of over-sensitivity to non-normal data. Levene's is based on the average of the absolute deviations of observations from the treatment mean. Levene's test is not easily available but it could be implemented by hand treatment of the data (Appendix P, Section P.2.3).

O'Brien's Test (O'Brien, 1979) is superior to Levene's test in certain mathematical aspects; however, it is almost unattainable, even in text books.

The *test of Bartlett* (1937) is standard for testing homogeneity of variance in software packages for environmental toxicology, and is described in Appendix P, Section P.2.3. It has the drawback of high sensitivity to data that are not normally distributed, especially skewed distributions. A set of data might be rejected because of an erroneous conclusion about homogeneity of variance.

Each of these tests starts with a null hypothesis that there are no differences in variance among treatments. If variances differ substantively, a subsequent ANOVA is invalid. These tests operate on the assumption that observations are normally distributed. Data based on proportions should not normally be put through these procedures (Appendix P, Section P.2.4).

7.3.2 Decisions after Testing Distribution of Data

Results which pass both the Shapiro-Wilk's and Levene's or Bartlett's tests should be analyzed using parametric methods, i.e., ANOVA. Data which fail to satisfy either test might be transformed to meet the requirements. The transformed data are tested for normality and homogeneity, and if the requirements are met, the data are analyzed using standard parametric methods. Transformation has complications and disadvantages, however, as described in Section 2.9.2.

If the original or transformed data do not satisfy either test for distribution of data, then analysis by nonparametric methods must be carried out (Figure 19).

The tests for normality and homogeneity of variance can be overly sensitive in some cases, while ANOVA and the multiple-comparison tests are rather robust towards minor non-conformity (Appendix P.2.4). Accordingly, if a data-set deviated mildly or moderately from normality or homogeneity of variance, an investigator might wish to consult a statistician about suitable procedures for analysis. It is recommended that both parametric and nonparametric analysis should be done, and reported. The more sensitive of the two analyses should provide the definitive estimate of toxicity.[54]

---

[54] There appears to be some support for flexibility in this respect. A group of statisticians and others who wrote an ecotoxicity analysis document for OECD (2004) had a

Results of Shapiro-Wilk's and O'Brien's (or Bartlett's) test should also be submitted, with a graph of raw results. The rationale for this is that parametric tests are often more powerful in detecting toxic effects, even with minor irregularities in the data.

7.4 Analysis of Variance

Analysis of variance (ANOVA) is carried out for parametric testing. It tests the *null hypothesis (H_0)* that there is no difference in the mean effect among treatments (concentrations). Most investigators will be familiar with analysis of variance, and it is available in most software packages for toxicology. It is further described in Appendix P.3. The ANOVA compares the variation among mean effects at the various treatments (concentrations), with the variation of effects for the replicates within concentrations. The ratio of the two is compared to critical values available in tables, to determine whether there is one or more significant difference(s) between treatments. If not, the analysis is at an end and the null hypothesis is accepted. If a difference is shown, analysis can proceed further by using a multiple-comparison test to identify differences.

Key Guidance

- *When data conform to normality and homogeneity of variance, the first step in parametric testing is an analysis of variance (ANOVA) to detect an overall difference among treatments. ANOVA compares the variation between concentrations with the background of variation within concentrations.*
- *If the ANOVA detected an overall difference, a multiple-comparison test would follow, to decide which concentration(s) caused effects different from the control. That determines the LOEC (lowest-observed-effect concentration). The next lower concentration is the NOEC (no-observed-effect concentration). Williams' test is recommended if there is an order of concentrations in the treatments, or Dunnett's test if there is no such order. Fisher's LSD test is recommended for pairwise comparison (each treatment with each other). Substitute tests are available.*
- *For nonparametric analysis of data which are ordered, Shirley's multiple-comparison test is recommended for comparing treatments with the control, although the method is not readily available. For pairwise comparison, the Jonckheere-Terpstra test should be used as a non-parametric analogue of ANOVA. If it rejects the null hypothesis, it should be followed by the Hayter-Stone test for a pairwise multiple comparison of treatment effects.*
- *For nonparametric analysis of unordered results, and comparison of treatments with the control, the Fligner-Wolfe test should be used to test the null hypothesis. If there is rejection, the multiple-range test of Nemenyi-Damico-Wolfe is recommended. If unavailable, alternatives are the Wilcoxon Rank Sum and Steel's Many-One Rank tests. For pairwise comparison, the Kruskal-Wallis test should be used for the null hypothesis. If it is rejected, the Critchlow-Fligner-Steel-Dwass multi-comparison test should follow; if that is not available, substitutes are listed.*

surprisingly relaxed view on the formal tests for normality and homogeneity. In the section on choosing between parametric and non-parametric methods, they state that "[a] visual inspection of the data may have indicated that the scatter is more or less symmetric and homogeneous ... In that case, one may analyze the data by the standard parametric methods based on normality." Further: "When the data appear to comply with the assumptions (after a visual inspection) of a particular parametric analysis, this is the obvious method to choose. The assumptions can be further checked at the end of the analysis (e.g., by examining the residuals ...). It may be noted that parametric analysis based on normal assumptions is reasonably robust to mild violations against the assumptions." "Formal tests exist as well ..., but it should be noted that mild violation of the assumptions is no reason for concern, and tests do not measure the degree of violation." For most investigators, it would be difficult to judge what constituted a "mild violation", in which case, advice should be sought from a statistician.

The ANOVA also provides what is called the *error variance* or *residual error term* for any subsequent multi-comparison tests (Section 7.5). Modern computer programs for ANOVA can process data with unequal replication and provide the correct residual error term for any subsequent multiple-comparison tests.

7.5 Multiple-comparison Tests

A multiple-comparison test is applied to determine which treatments elicit effects that are significantly different from effects in the control, and if desired, from each other. The various multi-comparison tests (sometimes called multi-range tests) make somewhat different comparisons. The investigator selects the appropriate one (Figure 19). Because this is hypothesis testing, none of the tests take into account the numerical value of the concentration, but two of them (Williams' and Shirley's) consider the mean effects in order of concentration and find the first one that is different from the control. Accordingly, when data are ordered, as in testing a series of concentrations, the first preference is to use Williams' (parametric) or Shirley's (non-parametric) test.[55]

Multi-comparison tests are discussed in the following subsections and further explained in Appendix P, Sections P.4 and P.5. Mathematical details are found in Newman (1995) or in standard statistical textbooks. Many of the important tests are offered in various computer packages.

7.5.1 Parametric Tests

Williams' test (Williams, 1972) is strongly recommended because it takes into account the order of concentrations according to their increase or decrease. This desirable feature, is appropriate for most toxicity tests. Williams' test compares effects at each concentration with the control, as is standard in many multi-concentration tests. The test statistics are compared, in order, with the critical value. The first statistic to exceed the critical value indicates a significant difference of that mean from the control. Because of its superior statistical power, Williams' test is appreciably more sensitive in estimating a lower LOEC than other available tests (Appendix P, Section P.4.1).

Williams' test assumes that the data within concentrations are normally distributed and homogeneous. There must also be a monotonic series of concentrations. If that is not the case, the means should be smoothed although that can reduce the sensitivity of the test. Smoothing might be available in newer toxicological software, otherwise the smoothing should be done by hand calculation. The test statistics are estimated by one of two simple formulae, depending on whether there are equal or unequal numbers of observations contributing to the mean values. The critical value, corresponding to the desired Type I error rate and the error degrees of freedom, is obtained from tables (Williams, 1972) if the data are "not too unbalanced" according to criteria described therein. Tables for unbalanced cases are available in Hochberg and Tamhane (1987). The test loses some power with unbalanced data, and OECD (2004) cites evidence that it should not be used for highly unbalanced results. Williams' test is further discussed in Appendix P, Section P.4.1, considered in detail in OECD (2004, appendix), and its procedures are given by Newman (1995).

Dunnett's test, like Williams', compares each group mean with the control, but is less powerful because it ignores the order of concentrations (Table P.3; Dunnett, 1955; 1964). If there is no implied order in samples, such as various sediments tested simultaneously at a single strength, then Dunnett's test can be used instead of Williams'. Dunnett's is given more prominence than Williams', in the computer programs used for environmental toxicology.

The basic formula for Dunnett's test is similar to that of the Student t-test. The common software packages for Dunnett's require an equal number of observations at each treatment. A series of modifications, which allow for unequal numbers, have been published, culminating in Dunnett and Tamhane (1998). Until a suitable modification becomes incorporated into the available software programs, investigators with such data could consult and use the published modification, or could use the Dunn-Sidak test described in the next paragraph.

---

[55] Statisticians might prefer approaches other than multiple-comparison tests, at least for parametric data. They might choose to initiate comparisons by using statements built into General or Generalized Linear Models (GLM, GLIM, see Section 6.5.2).

The ***Dunn-Sidak test*** could be substituted for Williams' or Dunnett's tests, if there were unequal numbers of replicates because of accidental loss or other cause. A ***Bonferroni adjustment of the t-test*** is frequently used, but is less powerful than the Dunn-Sidak test and has no particular advantage over it. Both tests are less powerful than Williams' and Dunnett's, in estimating the NOEC/LOEC.

An investigator might wish to compare differences among all pairs of locations in a multi-location survey. ***Fisher's Least Significant Difference,*** *(LSD)* related to the t-test, is recommended. It controls the family-wise Type I error rate and can deal with unequal replication, but is not common in computer packages designed for toxicology (Appendix P, Section P.4.4). The LSD is also intended for only a few of all the possible comparisons in a set of data, and those comparisons would have to be specified in advance. (The preceding limitation has general application for other multiple-comparison tests.) ***Tukey's test*** is similar, commonly available, can adapt to unequal sample sizes, but is not very sensitive (Appendix Table P.3). The ***Student-Newman-Keul test*** (SNK) is another alternative.

7.5.2 Nonparametric Tests

Nonparametric tests are strong tools for data that are not normally distributed. Generally, they tend to be less powerful than parametric tests if used on normally distributed data, in which case they might fail to detect a real effect of toxicity. Many of the commonly used nonparametric methods require at least four replicates; however, some do not (e.g., the Wilcoxon Rank Sum test).

It is recommended that nonparametric testing follow the same general sequence as is used in parametric testing. First, the null hypothesis of no difference in the treatments should be tested using methods that are analogous to an ANOVA. Only if the null hypothesis is rejected, should testing proceed to the multi-comparison tests.

Analogues of ANOVA. The ***Kruskal-Wallis Rank Sum test*** is sometimes provided in software packages, and can be used as the nonparametric equivalent of an ANOVA (Kruskal and Wallis, 1952; hereafter called the *Kruskal-Wallis test*). The ***Fligner-Wolfe test*** (Fligner and Wolfe, 1982) examines the null hypothesis that the treatment means are equal, with the customary alternative hypothesis that one or more treatment means differ from the control.

The ***Jonckheere-Terpstra test*** (Jonckheere, 1954) also tests the null hypothesis that the medians are equal, but the alternate hypothesis is that the *treatments are ordered*. It is suitable for data that strongly deviate from normality and homoscedasticity, and has very good power. This test has no problem handling unequal sample sizes, but failure to take into account the number of individuals in each subgroup can also be a disadvantage. Unfortunately, the method is not widely available as a computer program, and without that, it requires tedious manual calculations. However, a version that handles small sample sizes is available in the commercial software SAS and StatXact (OECD, 2004). The characteristics of the test are described in great detail in an appendix of OECD (2004).

These three tests start with the null hypothesis of no differences among the effects of treatments. As in parametric testing, if the null hypothesis of equality is accepted, then the statistical analysis would stop there, with a conclusion of no significant differences.

Multiple comparison. ***Shirley's test*** (Shirley, 1977) is recommended as first choice for comparing the treatment medians with a control median, *if there is an order in the magnitude of treatment and/or effect*. This is the nonparametric analogue of Williams' test, and it considers the order of concentrations. It requires five replicates but does not need equal replication. Unfortunately, it is not provided in most computer programs, nor is it easily available in printed form (Appendix P, Section P.5).

It is also possible to do a pairwise comparison (each treatment with each other treatment), if the treatments have an order (e.g., a series of concentrations). The Jonckheere-Terpstra test can be applied, and if the null hypothesis is rejected, analysis proceeds to the ***Hayter-Stone test*** for pairwise multiple comparison (Hayter and Stone, 1991). Unfortunately, here again, software for these test procedures is not readily available.

If the treatments have no order (e.g., locations in a general survey), a non-parametric analogue of

ANOVA should be applied first. Only if the null hypothesis is rejected (i.e., a difference exists somewhere among the treatments), should the analysis proceed to non-parametric multiple-comparison tests, as in the following text. This step of testing the null hypothesis is not necessarily stipulated in procedures outlined elsewhere, but it is recommended here as a suitably conservative approach. The procedure should eliminate or greatly reduce Type 1 errors, i.e., false conclusions of a difference. In statistical terminology, the multiple-comparison test is being "protected", by the initial test with an analogue of ANOVA. Since the multiple-comparison test is not run unless the preceding test rejects the null hypothesis, the multiple-comparison test is "protected" from finding a difference due to chance alone.

In a non-ordered situation, the ***Fligner-Wolfe test*** is recommended to test the null hypothesis of no difference from the control (Fligner and Wolfe, 1982; see Appendix P). If that is not available in suitable computer software, the Kruskal-Wallis test could be used. If the null hypothesis is rejected, the recommended first choice for comparison with the control is the ***Nemenyi-Damico-Wolfe test*** (Damico and Wolfe, 1987). This is suitable for a balanced design (i.e., equal numbers of replicates). A second choice is the ***Wilcoxon Rank Sum test*** which is generally available, and handles unequal replication. This is also known by other names such as *Wilcoxon signed rank test*, and often in Europe as the *Wilcoxon-Mann-Whitney test* or simply the *U test* (Appendix P, Section P.5.4). It is often used without the initial test of the null hypothesis, but carrying out that step is supported by Hollander and Wolfe (1999). A third choice that is commonly available in toxicological software, is ***Steel's Many-One Rank test*** (Steel, 1959), which requires equal replication.

If pairwise comparisons are desired for a non-ordered set of data, the Kruskal-Wallis test should be used to test the null hypothesis. If the hypothesis is rejected, the first choice of a follow-up test would be the ***Critchlow-Fligner-Steel-Dwass test***, also known as the *Critchlow-Fligner test* (Critchlow and Fligner, 1991). This is suitable for equal or unequal replication. If it is not available in suitable software, ***Steel's Pairwise test*** (Steel, 1960) should be used for balanced data. This test should not be confused with the earlier *Steel's Many-One Rank test* (Steel, 1959; see previous text.) For unbalanced sets of data, a somewhat unusual procedure could be followed. First the null hypothesis is tested with the Kruskal-Wallis test, and in the case of rejection, the same test is used for multiple comparisons, to find which treatment mean(s) differ from which others.

Edwards and Berry (1987) developed a multiple-comparison test which can be used in all situations, but unfortunately, is not readily available as software.

Dual-effect Tests

Dual-effect tests measure two different effects, usually mortality as a *quantal* component, plus a sublethal component such as weight of organisms or number of progeny, which is almost always *quantitative*. These categories of effect have been previously covered (see Sections 4, 6, and 7), but in dual-effect tests there are conceptual and statistical quandaries because the two effects frequently interact. For example, the weight of individuals which die during the test is lost from the assessment because it is not possible to determine those weights. Similarly, the death of an individual could obviously affect the number of young that it could have produced.

Choice among methods of analysis is partly governed by the philosophical/biological aspects of applying results to the real world, and partly by the practical aspects of particular toxicity tests. Separate sections follow for quantal and for two categories of sublethal effects.

8.1 The Quantal Component

Key Guidance

- *Quantal effects in sublethal or chronic tests should be analyzed by quantal techniques.*
- *For mortality that occurs during a test designed to measure chronic or sublethal quantitative effect, a separate analysis of mortality should usually be conducted by standard quantal methods such as probit regression.*
- *The LC25 might be estimated instead of the LC50, if an endpoint somewhat parallel to the sublethal IC25 is desired.*
- *For sublethal quantal effects such as egg fertilization, an ECp should be estimated by the usual quantal techniques, although quantitative analysis may be used if there are ≥ 100 observations per replicate.*
- *In dual-effect tests involving reproduction, it might be desired to analyze that effect in combination with mortality, using a "biomass" approach.*

The quantal part of a dual-effect test is usually mortality, and dealing with it is sometimes relatively straightforward. Investigators must not assume that because a toxicity test is chronic, mortality should be analyzed as an Inhibiting Concentration (ICp). Mortality is a quantal effect and should be analyzed by quantal techniques (Section 4). The collected data on mortality must still be considered quantal, even if they result from the cumulated effects of a variety of sublethal actions during a chronic exposure.

Estimating the LC25, to parallel the customary quantitative endpoint of IC25 could be done, although confidence limits would be wider than for LC50 (Figure 7). There cannot, however, be an extrapolation to the endpoint, so there *must* be an actual observed effect ≥25% in order to estimate an LC25. Maximum mortality could, however, be less than 50% [56].

Other quantal endpoints can be obtained in dual-effect tests such as success in fertilization of salmonid eggs, which should be analyzed using quantal methods (Section 6.1.1). Numerous quantal observations (≥100 for each replicate) can be analyzed by quantitative means (Section 6.1.1).

Sometimes mortality is intimately combined with reproductive effects, and it is appropriate to analyze the combined effects in a "biomass" approach (see Section 8.3).

[56] As noted in Section 4.5.3, some computer programs for estimating ECp will not analyze data unless there is an effect ≥50%, a safeguard to prevent estimates of EC50 from inadequate data. For EC25, that restriction would have to be circumvented, or another procedure used.

8.2 "Growth" as the Sublethal Component

Key Guidance

- *In a dual-effect test which measures attained size (so-called "growth"), it is often preferable to analyze that sublethal effect separately from any mortality, to estimate an independent endpoint, usually the ICp. For attained weight of salmonid fry, fathead minnow larvae, juvenile amphipods, or midge larvae, the separate analysis can be based on the average final weight of survivors in each replicate. Dead individuals would not contribute data for the endpoint based on weight. Nevertheless there might be a bias caused by interaction, say if "weak" individuals showed both smaller size and more rapid mortality; no method of dealing with that problem is evident.*
- *An alternative, the "biomass" approach, combines mortality with size by analyzing total weight of survivors, or total weight in a replicate divided by the number of organisms that started the test in that replicate. It can be used, if desired or prescribed. To some extent it simulates ecological success, and might yield stronger effects. This approach also has the potential bias from interaction of size and survival time.*
- *The "separate" and "biomass" approaches should not be mixed in half-measures.*
- *Mathematical techniques must be chosen with care. The "separate" approach could result in unbalanced numbers in replicates, limiting the suitable statistical methods. The "biomass" approach could produce measurements of zero in some replicates, leading to complications with variance.*

The sublethal component of dual-effect tests is usually quantitative, such as weight or number of young. Analysis is more straightforward for weight of organisms, and the choices are explained in this Section (8.2). These are often called "growth" tests, although the data would be better described as "attained weight" or "attained size". Usually there is a measurement of size at the end of the test, but none at the beginning, as would be required for a proper assessment of growth.

Choice of an approach for number of progeny is more complex (see Section 8.3).

8.2.1 Options for Measurement

In dual-effect tests dealing with attained weight (or attained length or other measurement of size), there should be thoughtful consideration of the sublethal *effect* that is to be analyzed and reported. The magnitude of the endpoint that emerged might be much higher or lower for certain effects or combinations of effects. Usually, the basic choice is whether to combine the sublethal measurements with mortality, or to attempt to keep them separate. For dual-effect tests, some method documents of Environment Canada specify which procedure to use, and the specification must be followed for departmental programs. If the choice were open to the investigator, it would be partly philosophical, depending on the investigator and the ecological applications of results. Nonetheless, the choice has definite implications for validity of mathematical procedures.

Whatever choice is made, there could be subtle unknown and undesired interactions in the test containers. For example, if some organisms died, that might allow more test solution, more space, and/or more food for the remaining organisms, possibly affecting growth or well-being of the survivors. Such possibilities should be considered when interpreting results. Although no statistical correction can be made for such interactions, their importance can be minimized by following the recommendations of Environment Canada on volume of test solution and other procedural matters.

Another potential difficulty which could lead to an over-sensitive sublethal test has been called *the scrawny/brawny interaction* by statistician B. Zajdlik. It could be quite possible that "weak" or "enfeebled" individuals might die first in a test, and would also be of smaller size. At a low concentration in which there was no mortality, such "scrawny"

individuals would survive and influence that concentration to show a representative, but relatively low mean weight. At a higher concentration, only the "brawny" would survive, biasing the mean weight upwards for that concentration. The net effect would be that the estimated endpoint for effect on weight would move downwards.

There are three basic options that have been used for measuring effects. The choices are illustrated in this section by the sublethal test on fathead minnows, which measures final weight attained by a group of larvae in a given replicate (EC, 1992b). The options represent different objectives; and are not equal from either a biological or statistical viewpoint. All of the options have minor or major imperfections or difficulties.

Option (1) Separate the sublethal effect from any mortality in the test, and analyze it separately, as far as that is possible. This means tabulating the sublethal measurements, only for organisms that survived until that measurement was made (at the end of the test). For fathead minnow larvae, the raw data would be the **average weight of surviving fish**. The total weight measured in each replicate at the end of a test, would be divided by *the number of larvae that survived* in that replicate. As indicated previously, that might lead to a "scrawny/brawny interaction" of unknown magnitude, for which there would be no remedy. If no fish survived in a replicate, there would be no measurement of weight, and no entry of data (in essence it would be a missing replicate). Mortality would be assessed by a separate analysis (Section 8.1).

Option (2) Partial allowance for mortality has sometimes been used in the following inconsistent fashion that is not recommended. If there were one or more living organisms in a replicate, then the weight would be estimated as the average weight of surviving fish, as in Option (1). If all 10 organisms died in a given replicate, zero would be entered as the sublethal measurement. By this method, zero is being used as the average weight of 10 dead larvae in the replicate, which is inconsistent. If all larvae died, zero weights would be used to represent them; if some larvae survived in a replicate, zero weights would not be used to represent the dead larvae.

Option (3) The "biomass" endpoint is a combination of sublethal effect and mortality. It can result in major differences among observations for different concentrations, and gives strong emphasis to the effect of the test material. In the fathead minnow test, the measurement analyzed would be the total weight of living fish in a replicate at the end of the test, divided by the *number of larvae that started* in the replicate. (If the same number of larvae had started in each replicate, exactly the same result would be obtained by analyzing the total weight of fish in each replicate, rather than the average.)[57] The **final biomass** is the measurement analyzed. If all fish died in a replicate, then a value of zero weight would be assigned, as in Option (2) [58]. This approach might also have the bias of incorporating the "scrawny/brawny interaction".

All three options have been used in Canada, and two of them have been recommended or at least suggested in methods published by Environment Canada.

Option (1) is the standard practice in Environment Canada's method for weight of fathead minnow larvae (EC, 1992b) [59]. It is also standard for analyzing the weights of salmonid fry in the early-life-stage test

---

[57] If any larvae had been accidentally lost or damaged during the exposure, they would be deducted from the initial number of larvae in that replicate.

[58] In effect, Option (3) assigns zero weight to dead larvae in any replicate, whether the replicate had complete or partial mortality. The method is therefore an extension of Option (2), but removes inconsistency. The procedure fits with its name: the "biomass" approach.

[59] The Environment Canada method for sublethal effects on fathead minnows instructs that a concentration should be excluded from the analysis if all the larvae died in all the replicates of that concentration (EC, 1992b). The instructions are not explicit on what should be done if all the larvae died in one replicate, but not in other replicates of the same concentration. The consistent action would be to leave the replicate with complete mortality out of the analysis. This action would result in an unbalanced set of data, and would require a method of statistical analysis that was appropriate for unbalanced replicates. Certainly the EC instructions do not require Option (2), which would have meant entering a zero for weight of a replicate which showed complete mortality of larvae.

(EC, 1998a), of midge larvae (EC, 1997a) of the freshwater amphipod *Hyalella azteca* (EC, 1997b), and of the polychaete worm *Polydora cornuta* (EC, 2001a) in sediment toxicity tests.

Option (2) does not appear to be standard for any published methods. Nevertheless it was commonly used for weight of midge larvae in tests of sediment toxicity, by some Canadian consulting companies, before publication of the method by Environment Canada (1997a).

Option (3) is common practice in certain test methods of the USEPA, in which weight or size is the measurement of effect. The ICPIN program (Norberg-King, 1993) instructs that the total weight of fathead minnow larvae in a replicate is to be divided by the number of larvae that started the test.

8.2.2 Conceptual Aspects of the Options

The three options of Section 8.2.1 have various positive and negative features.

Option (1) can certainly be justified on biological grounds, as it directly examines sublethal performance, and *only* sublethal performance, of all the test organisms which proceeded through a complete exposure. This approach seems to be rational, but it can produce anomalies. For example, in some long-term tests with amphipods, mortality is a more sensitive endpoint than growth. Test sediments might even have better nutritional quality than the control sediment, and produce better growth of amphipods (U. Borgmann, 2001, pers. comm., National Water Research Institute, Environment Canada, Burlington, Ontario). This kind of anomaly is remedied in test methods of Environment Canada, which require a separate analysis of mortality, with the most sensitive effect adopted to represent the test. Also, there is the possible bias of a "scrawny/brawny interaction" with an unknown magnitude.

Option (2) is a compromise, used historically, which cannot be justified from a conceptual viewpoint. As mentioned, the procedure was used unofficially in Canada for tests with midge larvae in sediment. Obviously, each larva had a finite weight at the beginning of a test, and assigning a final weight of zero to any of those larvae is not a rational representation. As an extreme example, if all the larvae died in a replicate, entering zero for weight would imply that there were larvae alive at the end of the test, and they had absolutely no weight. This certainly influences the distribution of measurements and moves the endpoint to a lower concentration, but the approach is internally inconsistent and undesirable.

The "biomass" Option (3) can be justified on ecological grounds, since it simulates the overall success of the species under the conditions of exposure. Ecological success is often measured in terms of total biomass or total number of individuals. Option (3) is likely to give a more extreme (steeper) dose-effect curve than Option (1), probably with a lower concentration as endpoint. However, the data of Option (3) have more variability, with reduced statistical sensitivity to balance the apparent increased biological effect (Zaleski *et al.*, 1997). Indeed, in tests with fathead minnows, Option (3) has produced estimates of toxicity that are no lower than those of Option (1) (Pickering *et al.*, 1996; WSDOE, 1998). Option (3) could be appropriate in long-term tests such as those for amphipods in sediment, which have mortality as a sensitive endpoint. Again, this option might contain the bias from a "scrawny/brawny interaction".

8.2.3 Statistical Aspects of the Options

The three options have potential statistical intricacies. In each, the numbers of individuals might be different in various replicates and concentrations, requiring more complex statistical procedures. Unbalanced numbers would not be a serious problem for point estimates based on regression.

Option (1) seems to have the least serious problems for analytical treatment. Unequal numbers in replicates could be compensated by standard methods in either regression or ANOVA. Some difficulties might arise. If there were wholesale mortality at high concentrations, those concentrations would be missing from the sublethal analysis. If growth was affected only near concentrations that eventually caused death, then sublethal observations in the upper part of the dose-effect curve would be missing or scanty, and the estimate of the sublethal endpoint could be inadequate or weak. Such a situation would be relatively uncommon, but could happen. The situation would be helped by designing in more concentrations with smaller differences between adjacent concentrations; Environment Canada suggests as many as 8 to 10 concentrations in dual-effect tests.

Option (2), a poor practice which was sometimes used in the past, has the previously mentioned problem of unbalanced treatment of dead organisms in those replicates with complete mortality, compared to replicates with partial mortality. At the least, that would mean unbalanced numbers in replicates, whereas analysis methods might have been designed for balanced numbers. Beyond that, any analysis would seem to be hopelessly compromised by having two categories of data.

Option (3) could have the common mathematical problem of unbalanced numbers in replicates and/or concentrations, which can be dealt with by appropriate statistical methods.

At the research level, a potentially superior approach was offered by Wang and Smith (2000). It differs from the previous options, but is statistically complex and not completely developed. The modelling includes *both* mortality and sublethal effects, and estimates an ICp based on both, complete with confidence limits. The authors concede that the fit of their model was not completely satisfactory. They indicate that "more complicated models" might be more suitable; apparently their already complicated statistical method is not an immediate solution to the difficulties mentioned in this section.

8.3 Number of Progeny as the Sublethal Component

Key Guidance

- *In a dual-effect toxicity test measuring mortality and number of progeny, assessing the combined effect in a "biomass" approach is a choice for analysis.*
- *The other legitimate approach, a separate analysis of the sublethal effect on reproduction (e.g., in* Ceriodaphnia*) is more complex than for tests which measure growth. This is because the number of progeny depends, in part, on the length of survival by the parents.*
- *A suitable approach can be based on an inspection-by-inspection tabulation of the average number of new progeny, per parent alive during that inspection period. The procedure deserves standardization, with a convenient computer package.*

If a dual-purpose test measures number of progeny ("reproduction") as the sublethal effect, there is another complexity which applies, in addition to those described in Section 8.2. The situation is illustrated by the test for reproduction of the water-flea, *Ceriodaphnia* (EC, 1992a), but also applies for reproduction of earthworms and springtails (Collembola, EC, 2004a,c). In the test with *Ceriodaphnia*, each parent daphnid starts the test in a separate container, and accordingly, represents a replicate at a given concentration. The number of young that it produces by the end of the test is the sublethal datum that is used in statistical analysis for the replicate. (In addition, the mortality of the parent daphnids is analyzed by quantal methods to estimate an endpoint such as LC50 or LC25.)

The EC test method bases the analysis and interpretation on this straightforward count of the actual number of young produced in each replicate, whether the parent survived or not, which is appropriate for the "biomass" concept.

8.3.1 Inter-relation of Mortality with Reproduction

If a parent daphnid dies before reproducing, the number of young for that replicate is zero. However, if a daphnid lives to reproduce, the observed number of progeny depends partly on the *duration* of parental life, since there would normally be repeated broods of young. Thus, the apparent clear-cut measurement of sublethal effect for the *Ceriodaphnia* test (number of progeny produced in a container during the exposure) actually has parental mortality integrated into it.

This particular kind of inter-mixing with mortality is not a factor in the sublethal endpoint for weight of fathead minnow larvae (Section 8.2). In tests with the fish, the mortality governed the number of larvae present at the end of the test. However, for the recommended Option (1), the criterion for a sublethal data-point was independent -- if a larva lived until the end of the test it contributed a weight to the observed sublethal data, but if it did not live, it did not contribute data on weight. Degree of mortality in a group did not affect the magnitude of the data-point (average weight), with the possible exception of the "scrawny/brawny interaction".

When number of progeny is the measured effect, there is an interaction with mortality, unlike the situation with weights of minnows. Parental mortality affects the number of young in the data-set, i.e., the magnitude of sublethal observations is shaped by the degree of mortality. In light of that, the three options for analyzing data (Section 8.2.1) can be considered for the test of reproduction in *Ceriodaphnia*. Current statistical analyses for reproduction of daphnids assume a normal distribution of the data, but they should be based on a Poisson distribution.

Option (3) is conceptually and mathematically suitable for analysis of data from the test with *Ceriodaphnia*, if the biomass approach is accepted as a suitable criterion of effect. Zeros, low numbers, and high numbers of progeny are entered into the analysis, disregarding the duration of parental survival. This method is, indeed, standard practice in Canadian and US tests of reproduction in *Ceriodaphnia*.

Option (2) was described in Section 8.2, and is not considered here for the reasons previously stated.

Option (1), which is not recommended, would encounter the additional difficulty described previously, that parental mortality cannot easily be separated from the sublethal effect, although such separation is the thrust of this option. If an adult dies before producing young, the zero progeny recorded for that replicate does not represent a sublethal effect on reproduction, but rather, it represents mortality. Similarly, if a parent died early, the low number of progeny would reflect that mortality, rather than a dampening of reproductive mechanisms.[60] A potential new approach to this difficulty is outlined in Section 8.3.2.

8.3.2 Analyzing Reproduction as a Separate Entity

A perceptive study of the problem of assessing number of progeny in toxicity tests with *Ceriodaphnia* was provided by Hamilton (1986). He documented a potentially favourable approach for option (1). This approach, of separating the sublethal effect from mortality, deserves to be assessed for future use. It is surprising that this has not been done already.

Hamilton (1986), used real data from a test with *Ceriodaphnia* to demonstrate the biases if numbers of progeny are based on either the initial number of adults, or the number surviving at the end. A possible solution is to tabulate the number of progeny produced by each live adult, at each of several inspection times. (The test usually lasts for seven days, and progeny are counted and removed daily.) The daily average-per-adult is calculated for all the replicates of a given concentration.

This approach is only valid if there is no correlation between mortality and production of young. If approaching mortality slowed down reproduction, the interaction could make this method inappropriate. Hamilton (1986) demonstrated that the correlation was negligible or absent. The parent *Ceriodaphnia* continued to reproduce at a normal rate until death, as far as could be detected by statistical means, and by convincing graphic comparisons. At the same time, Hamilton's examination of data indicated that the biomass approach reflected primarily the mortalities, not the reproductive rates.

---

[60] The situation for Option (1) would become even more untenable if there were a naive attempt to express the data for a given concentration as average number of progeny per adult. Any tactic for making that calculation runs into trouble. If one or more adults died part way through the test after producing some progeny, it would become difficult to calculate a realistic average without incorporating survival time. If the total number of progeny were divided by the number of parents that started the test, the average would be biased downwards (e.g., an adult that died on the first day of exposure without producing young, would still be used in the calculation as if it were a producing parent). If the number of progeny were divided by the number of parents surviving to the end, the average would be biased upwards, above a realistic value. (For example, an adult which died an hour before the final inspection would probably have produced its full component of progeny, but would not be used in calculating the average number per adult. As an ultimate absurdity, if all the adults produced their young, but died one hour before the final inspection, then a large number of young would be credited to zero adults, scarcely realistic.)

At the end of the test, the daily averages (of progeny per parent) were combined to produce a total average progeny per adult, at each concentration. Those base data represent a relatively unbiased estimate of reproductive performance. Hamilton (1986) demonstrated from adjusted data that the method would detect changes in reproductive performance, aside from any influence of mortality. To measure the variation at each concentration, Hamilton recommended bootstrap procedures.

The rationale for the approach appeared to be well thought out and well documented by Hamilton (1986). It is recommended here, as an approach which could be developed in the future, for dual-effect tests with reproduction as the sublethal effect, as in the test with *Ceriodaphnia*. This would be an "Option (1)" approach, separating number of progeny as an individual effect. The approach is similar to that used in human epidemiology, to investigate the expected time-to-death from, a given cause (say, heart attack), if the effects of competing causes of death were removed. It is also similar to procedures used in fisheries biology to remove the effect of "fishing mortality" so that natural characteristics of fish populations can be described (Ricker, 1958).

The precise procedures for this method of analysis need to be standardized and a convenient computer package developed. The method could apply to any dual-effect test in which a cumulative response was used for each animal, and early mortalities could occur. The preferred endpoint would be ICp. Hamilton (1986) recommended that in addition to a mathematical analysis, graphs of the number of young produced daily in each replicate, should be plotted to assess the separation of mortality and production of young.

This "Option (1)" or "separation" procedure could be an alternative to analyzing "biomass", which is used in the Environment Canada test with *Ceriodaphnia*. The two approaches are, however, identical if all of the adults survive to the end of the test.

8.4 Summary and Recommendations

It goes without saying, that tests following methods published by Environment Canada must use the prescribed analysis. In other situations, the choice of a suitable method and statistical analysis must be made by the investigator to meet the needs of the study.

There are two broad legitimate options for analysis and interpretation of dual-effect tests. The first approach ("Option 1") is to separate the sublethal effect from the other effect (usually lethality), and analyze it separately. This separation of effects might be more technically informative.

The second approach is to combine the two effects in a "biomass" type of analysis. This option might enhance the apparent toxic effect, and the results might better predict overall ecological effects in the real world. The biomass approach could be suitable for particular tests or purposes. The general application of this option does not, however, have much support among active Canadian investigators (Schroeder and Scroggins, 2001).

Approaches which partially combine two effects, should be avoided.

In dual-effect toxicity tests which measure attained size of organisms, Option (1) appears to be preferable. This option uses the average size of the surviving individuals, and allows "clean" statistical analyses. Such observations are amenable to most common statistical methods which allow for unequal numbers of individuals in replicates. Mortality and other effects should not be ignored, but suitably reported after analysis by quantal methods.

In dual-effect tests, which measure number of progeny, the biomass approach is one suitable alternative for analysis and interpretation. It would be desirable for Environment Canada to develop and standardize an alternative approach, which would separate off the sublethal component (reproduction) by means of average number of progeny per parent, tabulated for each inspection period and summed for the test (see description in Section 8.3.2).

Some Statistical Concepts and Tools

Most investigators will already have a foundation in statistics, and this guidance document does not attempt to provide that. However, some statistical terms relevant to toxicology are defined in the Glossary for convenience. In addition, some mathematical basics that are relevant to toxicity analyses are outlined at the beginning of this section. Later parts of this section cover some often-used mathematical procedures.

9.1 Normal and Binomial Distributions

Key Guidance

- *Normal and binomial distributions are fundamental characteristics of quantitative and quantal toxicity tests, respectively. For large numbers, and proportions near 0.5, binomial curves become similar to normal ones.*

Normal distributions are basic for much of the data in toxicity testing, as in most biological fields. Many statistical tests assume normality of the data, particularly for quantitative sublethal tests (Section 6). Similarly, the binomial distribution is basic for quantal data (Section 4). For many observations and proportions near 0.5, the binomial distribution becomes similar to the normal.

9.1.1 Normal Curves

Features of a normal distribution are described in the Glossary, and a visual representation is given in Figure 20, using the heights of people as an example.

The characteristic normal distribution in the upper panel of Figure 20 shows most measurements of people's heights clustering around the mean. There are fewer and fewer observations towards the outer limits of the range. The frequency histogram can be described by a standard "bell-shaped" normal curve. Not all bell-shaped curves are normal; to qualify, a distribution must satisfy a fairly complex formula (Zar, 1999). Standard tests ascertain whether a set of data meets the requirements (Section 7.3).

The lower panel of Figure 20 shows how the magnitude of the standard deviation (σ, sigma) governs the shape of a normal curve. In this panel, "*X*" represents the measured variable, with a mean of zero in this case. The vertical axis represents frequency "*f*"(or *probability)* of occurrence. A greater value for standard deviation makes the curve wider and squatter. A change in the value of the mean would shift the curve to left or right, but would not change its shape. Normal curves are always symmetrical, although an asymmetric or skewed distribution might result if one normal distribution were superimposed on another (i.e., combining two sets of data with different means.)

9.1.2 Binomial Distributions

Binomial distributions are very important in environmental toxicology because a large part of the data is "either/or" in nature. Many tests count the numbers of experimental organisms that are dead, out of the total number exposed. Such data can be described as *binomial*, *binary*, or *quantal* (see Glossary). Some histograms of binary data are shown in Figure 21.

A symmetrical distribution is seen in the left panel of Figure 21, when the probability is 0.5. If the probability of the event is reduced, the distribution becomes skewed, as in panels (b) and (c). The frequency becomes higher for bars on the left side of the histograms, notably for zero occurrences out of five trials (X = zero, or no deaths). Consequently, the frequency becomes lower for bars on the right side of the histograms, notably there is a disappearance of five occurrences out of five trials (X = 5, or death of all five organisms).

From Figure 21, it is easy to see that with larger sample sizes (say, ≥25), and with $p \approx 0.5$, the binomial distribution of trials (or organisms) would assume the general shape of a normal distribution (Figure 21). Many of the observations would cluster

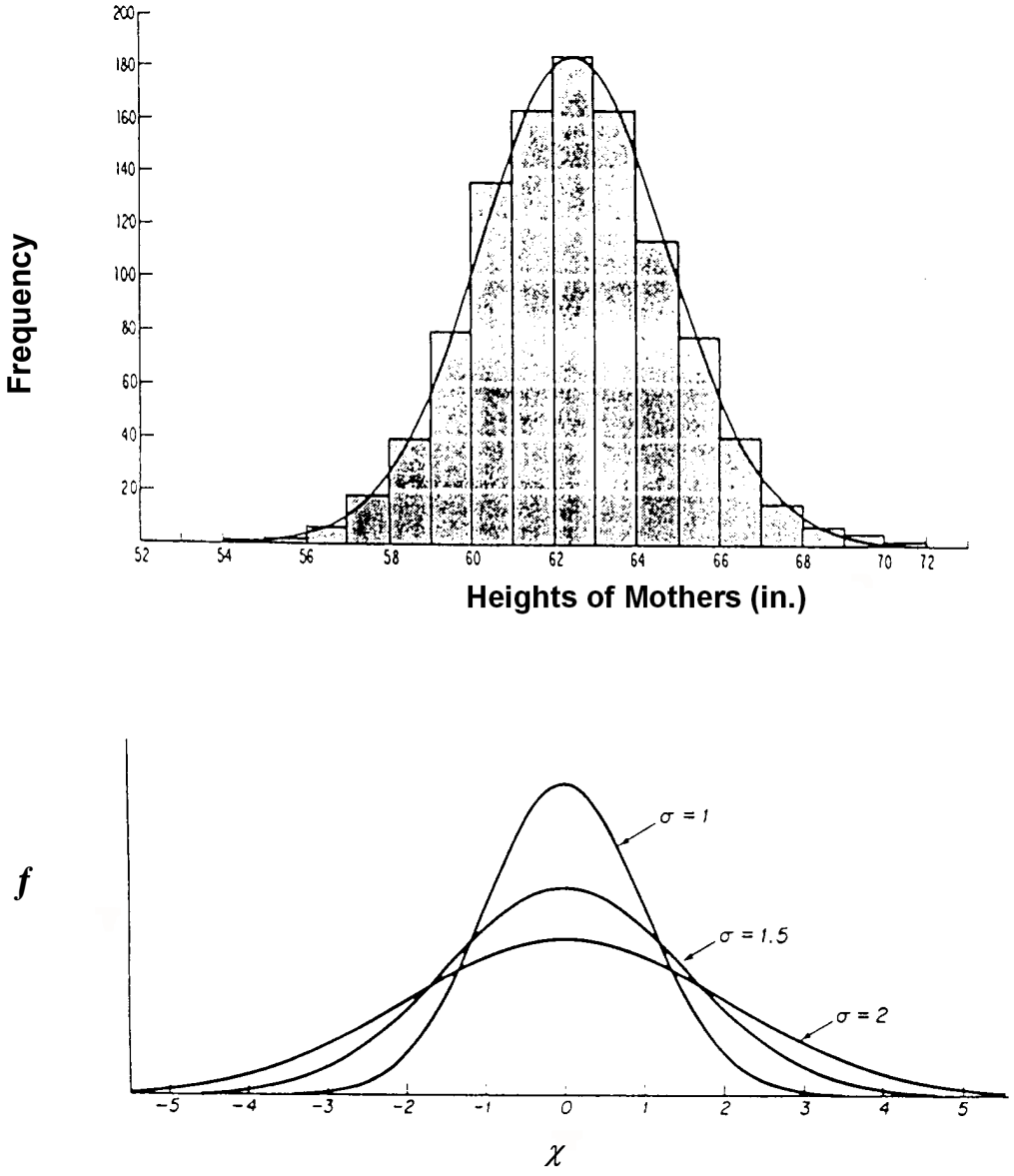


Figure 20 **Normal distributions.** The upper panel shows the distribution of heights of 1052 people, fitted with a standard "bell-shaped" normal curve. In the lower panel, "*X*" represents the measured variable, with a mean of zero; the vertical axis represents frequency *(f)*. The shape of the curve is governed by the magnitude of the standard deviation (σ, sigma). After Snedecor and Cochran (1980) and Zar (1974).

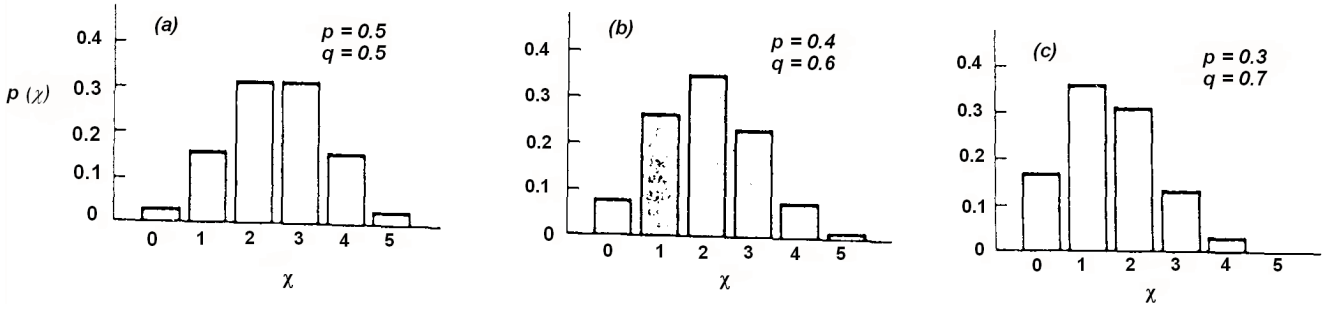


Figure 21 **Binomial distributions.** These are distributions for five trials of a binomial event (for example, "death" or "not death" among five waterfleas in a chamber). The probability of the event occurring ("death") is *p*, and the probability of it not occurring ("alive") is *q*. The horizontal axis (*X*) under each histogram represents zero occurrences, one occurrence, two occurrences, etc. out of five trials or replicates (i.e., no death, one death, etc. among the five organisms). The vertical axis is the frequency of those occurrences. In the left-hand panel (a), the probabilities are even for the event to occur or not, and a symmetrical distribution is seen. In panels (b) and (c) the probabilities of the event occurring are reduced ("death" is less likely), and the distributions are skewed. After Snedecor and Cochran (1980).

near the proportion 0.5, with fewer and fewer observations as proportions diverged further towards zero or 1.0. If *p* diverged appreciably from 0.5, the normal distribution would be a poorer approximation of the binomial one. Depending on the value of *p*, it might require hundreds of binary observations to achieve a distribution that was similar to the normal one. This characteristic is relevant to the assumptions of normality used to assess fertilization in the toxicity tests with echinoids and salmonids (EC, 1992f; 1998a).

9.2 Samples and Populations

Key Guidance

- *Toxicity tests always use a sample of test organisms, and random selection is essential if the sample is to represent the population in a holding chamber.*
- *There is seldom any attempt to establish whether a particular toxicity test is representative of the much larger free-living (wild) populations of organisms. However, most deliberate trials of "field validation" confirm that toxic levels determined in the laboratory are good predictors of harmful effects to natural (wild) communities.*

Investigators carry out toxicity tests on a *sample* of organisms. They might take a sample from a holding chamber containing large numbers of the organisms. All the organisms in the tank might be regarded as a *population*. The investigator assumes that the sample is typical of the organisms in the tank, which is why some process for random selection of the samples is important.

The endpoint of the test, and its statistical descriptions, always characterize the *sample*.

Statistical tests and descriptions will take into account the size of the sample, and the variation in the observations, as a part of producing their estimate. A large sample is likely to produce a more precise endpoint. Hence there is usually competition in designing the experiment, between a desire for a larger sample to obtain more precision, and a desire for smaller samples to reduce the size of apparatus, amounts of test substrate, and time required to check effects.

It is usually a reasonable assumption that the endpoint for the sample also represents the population. However, if the investigator did a poor job of sampling (say, taking all large organisms), any statistical findings would apply to the sample, but might not apply to the population in the holding tank.

At a broader level, there is an implicit assumption, not dealt with here, that the organisms in the holding chamber, and the sample endpoint, represent a much larger population, such as all the wild organisms of the species tested. Such an assumption is seldom tested for a given set of laboratory toxicity tests, and that must be acknowledged by users of the test data. Therefore, it is essential to present information on the sample of organisms tested, such as genetic background, rearing history, and size. These items are required in methods documents published by Environment Canada.

There is, however, a large body of information on "field validation" of toxicity tests in the laboratory. Field work around some Canadian pulp mills showed that effects in nature agreed with expectations from laboratory tests (Scroggins *et al.,* 2002), and the laboratory assessments are also useful for predicting effects of metal mining (Sprague, 1997). There have been an appreciable number of deliberate field research programs to associate effects in natural (wild) aquatic communities with results of laboratory tests, and also similar experiments using controlled communities (mesocosms). Environment Canada provided a major review of that research, and concluded that in most cases, the laboratory tests were good predictors of effects in natural habitats (EC, 1999a).

Further relevant particulars can be found in the Glossary under *sample*, *population*, *sampling unit*, *experimental unit*, *treatment*, *replicate*, *random sampling*, *sampling error*, and *precision*.

9.3 Statistical Versus Biological Significance

Key Guidance

- *There are relatively few attempts to define the degrees of toxic effect that are biologically significant. Ideally, in hypothesis testing, such a level should be defined before the toxicity test. The test and its statistical analysis could then be suitably designed to assess biological significance. The test result would be that an adverse biological effect had been (or had not been) observed with 95% certainty.*
- *At present, statistical significance of effects is generally substituted for biological significance by default, but they do not necessarily correspond without appropriate design.*

Throughout this document and throughout environmental toxicology, reference is made to the statistical significance of results. This is particularly true for hypothesis testing. It is well-nigh universal to select as a criterion, 5% probability that any difference would occur by chance. If an observed difference is large enough that it would occur by chance only once out of 20 times (or fewer), the difference is considered significant. This level of significance means that if 20 toxicity tests were done with a harmless chemical, results of one test would be expected to show a significant difference from the control (a Type I error, concluding a difference where none really existed, Section 7.2.2).

The deficiency of the general approach is that biological significance is seldom defined, and so that concept cannot be designed into the toxicity test. When the test is complete, biological and statistical significance need not have any particular relationship to one another. Section 7.1.2 mentioned that the statistical no-effect concentration was associated with sublethal biological effects that averaged 14% "worse" than the control, and ranged as high as 38% (Crane and Newman, 2000).

The proper approach would be for the biologist or toxicologist to *start the process* by deciding what was an ecologically significant effect in a particular

situation (survival, growth, reproductive rate, etc.) and convey information to the statistician. The statistician, in turn, would build that degree of effect into the hypothesis test, and inform the toxicologist about the requirements for numbers of samples, replicates, or individual organisms, given a certain amount of variation. After analyzing test results, the conclusion would be that an ecologically significant effect had (or had not) been demonstrated with 95% certainty. (This assumes that *beta* had been set at 0.05.)

A rare and laudable Canadian judgement on biological significance has been made in the Disposal at Sea Program. The criteria for meaningful biological difference have been set at 20–30% divergence from the control in certain tests of sediment (Porebski and Osborne, 1998; Zajdlik *et al.*, 2000; see Section 7.2.5). (The findings must also be statistically significant, of course.)

This question of judgement is prominent in another major approach of environmental toxicology, the point estimate (Section 6). The endpoint, ICp, can have any value of *p* selected by the investigator. The IC25, for 25% reduction in performance (compared to the control) has gained general approval as an endpoint that has reasonable ecological meaning (Section 6.2.4).

The decision on what is a meaningful ecological effect has to be derived from biological criteria and the judgement of investigators. The selected degree of effect might vary with the type of effect. Perhaps a decrease of 50% in the number of eggs might not be considered of any ultimate ecological importance, but a 10% decrease in growth rate of individuals might be considered of major consequence.

Lacking initial decisions on biological significance, the potential disparity with statistical significance can go either way. A statistically significant effect might be a very small effect, of no biological concern. However, an effect that was not statistically significant might be a large effect, of major biological concern, which could happen in a test showing great internal variability. This latter conflict is perhaps of greater practical importance. An investigator is torn between the statistical results, and a responsibility to point out a major biological effect. One thing to avoid is lapsing into phraseology of first-draft master's theses, that "although not statistically significant, the large change in the ... indicates that ...". In fact, in a case like this, the investigator has not shown that there was *any* change from the control.

Paine (2002) gives an excellent description of the overall conflicts in relation to programs for monitoring environmental effects, under the current approach for experimental design.

> "Environmentally significant effects may not be statistically significant, and statistically significant effects may not be environmentally significant. Environmentally significant effect sizes are difficult to define, because they depend on environmental, sociological, political and economic issues and values. Consequently, we often treat environmental and statistical significance as equivalent, implicitly or by our actions. Legal, regulatory, and management discussions and decisions are often based on the statistical significance of results or effects. More generally, journal articles and consultant or government reports often provide only the statistical significance of effects (e.g., "fish fecundity was significantly lower in the Impact area than in the Control area"), ..."

Paine (2002) makes three recommendations for dealing with the sometimes opposing tugs of statistical and biological significance. Certainly the first two should be followed by investigators reporting their work. The second one is the essence of the argument given previously.

(1) Report the magnitude of effects and the confidence limits, not just whether the effects were statistically significant.

(2) "Make an effort to define environmentally significant effects, however difficult that may be."

Paine's third recommendation, to give up the "obsessive focus on statistical significance", would be best satisfied by designing toxicity tests so that the statistical result had direct meaning for biological effect.

9.4 Inverse Regression

Investigators should be aware that in the usual kind of environmental toxicity test, there is a complex statistical problem in estimating the endpoint and its confidence limits. The complexity is handled by the statistical treatment, so the investigator does not have to take any remedial action. However, the complication explains why specific statistical procedures must be used, and why confidence limits are often asymmetric.

Key Guidance

- *Concentrations start as the independent variable in a toxicity test. Variation in the test is measured in terms of biological effect, the dependent variable. The estimate of endpoint and confidence limits is, however, inverted into terms of concentration. This entails statistical complexities.*
- *The investigator is largely unaware of the complexity during the statistical analysis, but this explains why particular analytical programs must be used for toxicity data, and why confidence limits can be asymmetric.*
- *The inversion does not apply to endpoints based on time, such as the ET50, since observations and calculations are based on variation along the time scale.*

Investigators usually fix concentrations when they set up a toxicity test, making concentration the independent variable. The degree of biological effect seen in the organisms is measured as the dependent variable. This creates a fundamental conflict between the design of tests and the desired endpoints. Simply put, the concentrations end up being treated as if they were the dependent variable. Determination of the endpoint is inverted, in order to estimate the concentration necessary to cause a fixed level of biological effect specified by the investigator, i.e., such items as the EC50, IC25, and their confidence limits. The inversion entails statistical complexities in the programs used to analyze the data.

The test concentrations being initially fixed, are presumed to be without variation. The experimental observations on degree of effect are subject to experimental variation about the true effect. If a linear relation between the two is calculated, its variability continues to be in terms of the measured biological effect. That linear relationship, with its variation along the effects axis, is used to predict endpoints and confidence limits along the other axis, i.e., the concentration axis. For example, a median effective *concentration* and the *concentrations* marking its confidence limits, would be estimated from the fitted line and its variation in *effects* (see Figure 7).

The conflict is less evident in hypothesis testing. The estimate of an endpoint is straightforward, because it uses the observed variation in effect to determine which treatment causes an effect that is significantly different from the effect in the control. However, an inverted estimate takes over for confidence limits, which are derived from variation in effect, but calculated in terms of concentration.

The reciprocal switching of variables between dependent and independent can be described as *inverse estimation of endpoints* and confidence limits. As mentioned, the inversion procedures are designed into statistical programs, so an investigator is not reminded of them. They remain, nonetheless, a complexity in the statistical procedures of most tests in environmental toxicity.

One common effect of the inverse estimation is shown by the example for a hypothetical quantal toxicity test in Figure 7. At any given concentration, the confidence limits are vertically symmetrical because they were calculated in terms of the *observed effects* at the fixed concentrations. However, the limits of the EC50 are along the horizontal axis for concentration, and will usually be asymmetrical because of the inversion in calculations. The asymmetry is noticeable if a ruler is laid horizontally at the 50% effect-level in Figure 7, or any other effect-level. Asymmetry is particularly evident near the ends of the probit line, where one or both of the limits can sometimes become very large or near-infinite.

The inverted estimates apply when any technique of regression, whether linear or nonlinear, is applied to the usual toxicity tests. Ordinary statistical packages (non-toxicological ones) do not provide a standard option for dealing with this, for example in estimation of confidence limits. This is one of the reasons why a specially designed program of probit regression must be used to estimate an EC50 rather than a simple line-fitting procedure based on least squares. Although a formula was provided by Nyholm *et al. (*1992) for estimating confidence limits about a toxicological endpoint derived by ordinary linear regression, it does not yet seem to have been incorporated into North American software packages for environmental toxicology. Formulae for the same general purpose are provided in Draper and Smith (1981) and in some other textbooks on regression.

The inverse regression does not apply to quantal tests for estimating the time to 50% effect (ET50 or LT50, see Section 5). Both endpoint and confidence limits are estimated in terms of the dependent variable, time. The direct approach is statistically clean, adding to the other benefits of using ET50 as an endpoint.

The other general approach to the problem of inverse estimation is to *reparameterize* the equation relating effect to concentration (see Section 6.5.12). Environment Canada has done that in recent test methods for soil toxicity (EC, 2004a–c; Sections 6.5.7 and 6.5.8).

9.5 Significant Differences Between EC50s

Key Guidance

- *Significant differences between two quantal endpoints (EC50s) can be assessed from their confidence limits. The simple comparison is similar to standard error of the difference.*
- *A superior mathematical method for two EC50s seems feasible.*
- *To test for differences among several EC50s, conventional ANOVA could be used for the unusual situation in which replicates were available.*
- *It seems possible that a dedicated mathematical formula could be developed to detect whether a significant difference existed among several EC50s, but like an ANOVA it would not single out which EC50(s) differed.*

Significant differences between endpoints may be calculated without recourse to ad hoc methods, when the raw data are available. These methods are, however, beyond the scope of the present document. This section describes ad hoc methods which can be used when the raw data are not available.

9.5.1 Pairs of EC50s

Some ad hoc methods are available for comparing two quantal endpoints for statistically significant difference.

No overlap of confidence limits. It is convenient that for results of quantal tests, significant differences between some pairs of EC50s may be determined by inspection of their confidence limits. If the confidence limits do not overlap, the EC50s are different, and may be declared so without further statistical testing. However, if the confidence limits overlap, this does not tell anything about significant difference.

Litchfield-Wilcoxon method. The method of Litchfield and Wilcoxon (1949) can be used to distinguish two EC50s. It is parallel to a recognized mathematical technique, standard error of the difference between means (Zar 1974, p. 105–106), although most statistical/toxicological texts do not cover this explicitly. The Litchfield and Wilcoxon (1949) method is analogous to the procedure for obtaining a single *pooled estimate of variance* from the variances of two distributions (Snedecor and Cochran, 1980). Finney (1971, p. 110–111) shows a parallel example, for obtaining a single variance for relative potency from the sum of two variances of a pair of substances. The method is questioned by Hodson *et al.* (1977) but is a standard one in pharmacology. Application to environmental toxicology is described by Sprague and Fogels (1977). The method has been used for some decades, and seems valid for pairs of tests which have similar distributions of data.

This approximate procedure is carried out as shown in Equation 7. The method might be used with

caution [61] until a superior mathematical method is developed and published.

$$f_{1,2} = antilog \sqrt{(log f_1)^2 + (log f_2)^2}$$

[*Equation 7*]

To compare two EC50s with overlapping confidence limits, the statistic $f_{1,2}$ is calculated according to Equation 7. There would be a significant difference if the ratio (greater EC50) ÷ (lesser EC50) exceeded the statistic $f_{1,2}$. The value f_1 is simply the ratio between the confidence limit and the EC50 for a given test, and may be calculated as [(upper confidence limit)/(EC50)] + [(EC50)/(lower confidence limit)], then dividing this sum by 2. For the other EC50, f_2 is calculated in the same way. A short computer program for this has been available from Environment Canada in North Vancouver [62].

The primary use of Equation 7 should probably be to determine whether two EC50s are *not* different, thus avoiding over-interpretation of variation which has not been shown to be real [63].

Caution should be raised that a conclusion of significant difference would apply only to the two particular endpoints that were compared, and might not prevail if additional tests were done. For example, if Equation 7 showed that EC50s for copper were significantly different for two species of crustaceans, it would not necessarily mean that the species were different in their tolerance, just that these two particular endpoints were different. Also, the biological significance of differences and possible causes should be considered. For example, variation in results from different labs or different times might lead to statistically significant differences, but the biological meaning of the difference might be in the realm of unexplained variation, and should be considered in that light.

Zajdlik's ad hoc method no. 1

A mathematical method for comparing two EC50s can be based on the two-sample Z-test, which is explained in most statistical texts (e.g., Zar, 1974, p. 105–106). The general method was suggested for use in comparing two EC50s by Hubert (1992). It could be a useful method once the steps in the procedure have been described by Zajdlik (in prep.). Equation 8 gives the formula for calculations.

$$Z = \frac{\log EC50_1 - \log EC50_2}{\sqrt{(\sigma^2_{(\log EC501)} + \sigma^2_{(\log EC502)})}}$$

[*Equation 8*]

Sigma (σ) represents the standard error, i.e., the standard errors of the first and second log(EC50)s.

---

[61] Equation 9.1 parallels that for Standard Error of the Difference, in which SE_{DIFF} equals the square root of the sum of (SE squared for the first item) plus (SE squared for the second item). Use of this method in environmental toxicology might sometimes be forcing it beyond its intended statistical foundation. In pharmacology, the classical procedures tested a drug of unknown potency against a standard of known potency. Testing for significant difference in potencies, in the manner of Equation 9.1, required the same slope for the dose-effect relationships of the two materials. In toxicity testing, the "slopes" of the effect-distributions might not be the same, so the validity of using this method is in question. It is likely that if f_1 and f_2 are similar, i.e., the confidence limits are of similar magnitude on a logarithmic scale, relative to their EC50s, then this procedure for testing significant difference would be reasonable. If the ad hoc procedure of Zajdlik becomes available, it would be the preferred method, in the absence of the raw data.

[62] Toxicology Program, Environment Canada, Pacific Environmental Science Centre, 2645 Dollarton Hwy, North Vancouver, B.C., V7H 1V2.

[63] Some examples might assist. For the comparisons of EC50s in the table, all the 95% confidence limits were arbitrarily fixed at EC50 * 1.5 and EC50/1.5. Thus $f_1 = f_2 = 1.5$, and $f_{1,2}$ will always be calculated as 1.77.

Higher EC50 (limits)	Lower EC50 (limits)	EC50 ratio	$f_{1,2}$	Different?
20 (13.3, 30)	8 (5.3, 12)	No overlap	Not tested	Yes
as above	11 (7.3, 16.3)	1.82	1.77	Yes, just
as above	12 (8, 18)	1.66	1.77	Not quite
as above	15 (10, 22.5)	1.33	1.77	No

The procedure for obtaining σ (and hence σ^2) remains to be defined at the time of writing.

If |Z| > 1.96, then the two EC50s are significantly different at the 95% level of significance for a two-tailed test, when the question posed is whether $EC50_1$ differs from $EC50_2$, greater or smaller. For a one-tailed test, in which the question is whether $EC50_1$ is statistically greater than $EC50_2$, significance would be established if |Z| is greater than 1.645.

Other approaches. There have been other thoughts on this topic as it relates to environmental toxicology. Villeneuve *et al.* (2000) address relative potency which is the same general question as determining significant differences between EC50s. They acknowledge that estimates of relative potency (ratio of two EC50s) are valid only when the dose-effect curves are parallel and show the same maximum achievable effect. The requirement for parallelism is less important in toxicity tests.

Villeneuve *et al.* (2000) offer a framework for analysis which would require further development for use in environmental toxicology. They outline a method using multiple-point estimates over a range of effects from EC20 to EC80, to determine relative potency ranges. Villeneuve *et al.* offer a dichotomous "framework" for decisions on deriving and applying estimates of relative potency; however, they do not offer a particular mathematical technique for treating the toxicity data. Transformation to a straight line is suggested, if possible, by using log of dose and probit, logit, or logistic tools. Subsequent linear regression was used but they refer the reader to several "generalized linear models and other nonlinear regression techniques" that are in the literature.

9.5.2 Comparing Multiple EC50s

A pairwise test, such as those shown by Equations 7 and 8, must not be repeated between all possible pairs in a list of EC50s, because it will likely cause a false positive (Type I) error. If the 5% level of significance had been adopted, the repeated testing would be expected to show significant differences because of chance alone, in one of 20 comparisons.

The problem is parallel to that for repeated use of a t-test in situations where an analysis of variance would be appropriate.

In environmental toxicology, investigators have seldom tested for differences among a series of EC50s, probably due to the lack of a convenient test method or package. Testing by conventional analysis of variance would be valid if replicate EC50s were available, but that would not be the case in most testing programs.

The method described by Equation 9 might be used by investigators if it were proven valid and its procedures were described. Equation 9 remains tentative at the time of writing but might be developed (Zajdlik, in prep.). The method is based on the chi-square test and would tell whether or not a significant difference existed among an array of more than two EC50s. As in an ANOVA, use of Equation 9 would not distinguish which endpoint differed from which others.

$$\chi^2 = \sum_{i=1}^{k} w_i \left(\log(EC50_i) - \left(\frac{\sum_{i=1}^{k} w_i \log(EC50_i)}{\sum_{i=1}^{k} w_i} \right) \right)^2$$

[*Equation* 9]

For each *EC50,* $w = (1/SE_{\log(EC50)})^2$, i.e., the square of the reciprocal of the standard error (SE) of the EC50 on a logarithmic basis. Some of the steps in making the calculation are given in Appendix Q, with an example calculation. A spreadsheet or simple computer program could be used for ease of calculation.

The calculated chi-square (χ^2) is compared with table values for one less than the number of EC50s, for the selected probability value, usually p = 0.05. If the calculated value is greater than the table value, there is one or more significant difference(s) among the EC50s.

To determine which endpoint(s) differed from which others, a multiple-comparison test would be required, but a suitable one has not yet been documented.

9.6 Significant Differences Between ICps

Key Guidance

- *Pairs of ICps may be compared by a method derived from the two-sample Z-test.*
- *If several ICps were replicated, significant differences could be assessed by standard analysis of variance and multiple-comparison tests.*
- *Without replication, there is no method currently available for testing differences among several ICps.*

Section 9.5 outlines methods for testing significant differences between and among the endpoints of lethal and other quantal tests. Parallel methods can be used for assessing quantitative endpoints. Some procedures for pairs of endpoints are established, but methods for comparing several endpoints await further development.

The most commonly used quantitative endpoint in North America is the IC25. The following methods are valid for any value of p, so the general term ICp is used. It is understood that comparisons must be made only for the same values of p, i.e., IC20 with IC20, IC25 with IC25, etc.

9.6.1 Pairs of ICps

No overlap. If the confidence limits of ICps do not overlap, they can be declared significantly different without further testing. If the limits overlap, this signals nothing about significant difference. The principle is the same as in comparison of two EC50s (Section 9.5.1).

Litchfield-Wilcoxon method. The Litchfield-Wilcoxon method (Section 9.5.1) uses a combination of confidence limits of two EC50s to judge significant differences. Although it might seem easy to extend the method to ICps, statisticians concur that it is not suitable for this purpose.

Zajdlik's ad hoc method no. 2. This is similar to the ad hoc method shown in Section 9.5.1. The method also originates from the two-sample Z-test, described in most statistical texts (e.g., Zar, 1974, p. 105–106). Equation 10 gives the formula, but the steps in calculation await description (Zajdlik, in prep.). The mathematical manipulations are simple enough as indicated by Equation 10, and involve only the logarithmic values of the ICps and their standard error. This method assumes that ICps are normally distributed, or rather, for environmental toxicity tests, that the logarithmic ICps are normally distributed.

$$Z = \frac{\log ICp_1 - \log ICp_2}{\sqrt{(\log SE_{ICp1})^2 + (\log SE_{ICp2})^2}}$$

[*Equation 10*]

If |Z| > 1.96 (i.e., greater than the critical value of Z) then the two ICps are significantly different at the 95% level of significance, for a two-tailed test. In the more usual case, it would be obvious to the investigator that one of the ICps was numerically larger than the other. A one-tailed test would be appropriate (is ICp_1 statistically no greater than ICp_2?). Statistical difference would be established if |Z| were greater than the critical value 1.645.

In Equation 10, SE represents the standard error of a logarithmic ICp. The SE for each of the ICps is calculated as shown in Equation 11 (Zajdlik, in prep.).

$$SE_{(\log ICp1)} = e^{\left(\ln(\log UCL - \log ICp_1) + \ln(\log ICp_1 - \log LCL) - 2\ln(z_{1-\alpha/2}) \right)} \div 2$$

[*Equation* 11]

For Equation 11:

UCL is the upper confidence limit of the ICp, for (1 - **α**)% (customarily 95%),

LCL is the lower confidence limit of the ICp, for (1-**α**)% (customarily 95%), and

Z refers to the normal quantile for (1 - **α**)%. The quantile is *UCL* (95%) minus *LCL* (95%) = 1.96, and that numerical value is substituted in the formula.

The approach is based on the confidence intervals which were calculated at the same time as the ICps. The method is appropriate, whether the ICp and limits were derived by regression or by interpolation and bootstrapping. The procedure in Equation 11 uses the upper and lower confidence limits to provide two estimates of the variance for a given ICp, in this case ICp_1, the first ICp. The geometric average of those estimates is used to obtain a single value for the first SE. The SE of the second ICp would be estimated in the same way.

The natural logarithms in Equation 11 are part of the process of obtaining the geometric average of the upper and lower confidence limits. The logarithms (base 10) represent calculations for test exposures derived from a logarithmic series of concentrations (Section 2.3).

In some methods of calculating ICp, notably in the better regression methods, the standard error would be part of the output of analysis. The investigator could use that value without resorting to Equation 11, and proceed to the comparison by Equation 10.

An example of calculating an SE by Equation 11 can be given, using arbitrary choices of ICp = 10 mg/L and confidence limits of 6 and 16 mg/L. On the right side of Equation 11, the part within the parentheses becomes (omitting some digits):

[ln(log 16 - log 10) + ln(log 10 - log 6)
- 2(ln 1.96)]/2

[-1.5890... -1.5057... -1.3458...] / 2 = -2.2203...

Taking that as an exponent would yield for the equation:

$SE_{(\log ICp)} = 0.108571336$

ICp + SE would be calculated as:

log ICP + $SE_{(\log ICp)}$ = 1 + 0.10857... = 1.10857, or as the antilog, 12.8 mg/L

ICp - SE would be calculated as:

log ICP - $SE_{(\log ICp)}$ = 1 - 0.10857... = 0.89142, or as the antilog, 7.79 mg/L

9.6.2 Comparing Multiple ICps

If there is true replication of ICps in several sets of tests, differences among the sets may be tested by the standard methods of ANOVA, followed by a multiple-comparison test if desired. However, if there is a series of unreplicated ICps, there does not appear to be any method in use, for establishing whether there is/are significant difference(s) among those endpoints. A pairwise test, such as that shown by Equation 10, must not be repeated between all possible pairs in a list of EC50s, because that could result in a false positive (Type I) error.

The method shown in Section 9.5.2, for comparing several EC50s would seem to lend itself to comparing ICps, and might see future development for that purpose (Zajdlik, in prep.) Alternatively, efforts might be focused on obtaining the raw data from the tests, and applying more sophisticated techniques.

Dealing with "Difficult" Results

Toxicity tests can produce data in a variety of patterns that are difficult to deal with. Some difficulties are addressed here, mostly for sublethal tests, but not all have agreed solutions.

10.1 Variation

Key Guidance

- *A high level of variability in effects should not systematically influence an ICp upwards or downwards, but wider confidence limits will indicate less reliability of the estimate. In hypothesis testing, higher variability moves the NOEC/LOEC towards higher concentrations.*

If a linear model is used to estimate an endpoint, high variability might not change the endpoint, although it will result in wider confidence limits. Those limits will be reported, so reliability of the endpoint will be appropriately apparent to all users of the information.

If hypothesis testing is used to analyze data from a test, high variability will make the test less sensitive. The endpoint will be a higher concentration, a deficiency of the hypothesis approach that is at the base of discussions in Sections 7.1, and 7.2.2 to 7.2.5.

Once a toxicity test has been completed, the variability cannot be changed. The only way of minimizing the effects is by choosing the most appropriate and effective methods of statistical analysis. If similar new tests were being done, the most likely remedy at the design stage would be to increase the sample sizes, or sometimes, to refine the statistical experimental design, and remove or reduce the procedural sources of variation.

10.2 Outliers

From time to time, test results will include an *outlier*, a measurement that does not seem to fit the other values from the test. An investigator would probably first notice an outlier by inspecting tabulations, or by plotting the distribution of data, which is one reason an initial hand plot is emphasized in this document.

There is no mathematical or judicial procedure that can magically and definitively separate an error from inherent variation. Error and variation could be similar in magnitude, and the investigator must not yield to temptation, by arbitrarily discarding a point which does not seem to fit a presumed distribution. On the other hand, a divergent point should not be blindly processed in the usual manner -- it might indeed be erroneous, and have a bad influence on technical interpretations.

Key Guidance

- *If an apparent outlier is noticed, it must not be removed from the analysis without a strong reason.*
- *If there is an outlier, all test records should be scrutinized for human error. Holding and testing procedures should be reviewed as possible causes of a changed biological response. Alternative models for analysis should be considered, perhaps a simple transformation of data.*
- *The investigator should also apply suitable mathematical tests for evaluating outliers (as described in the text). However, the tests have deficiencies for use in toxicology, and their conclusions should be tempered by inspection of total variation in an experiment.*
- *Anomalies should be reported, along with any tests done on them, and conclusions about their status.*
- *Generally, overall analyses of test results should be done with and without the outlier, and both should be reported, indicating which is considered definitive, and why.*

The problem of outliers can be approached in a reasonable manner. There are three successive steps in dealing with suspect observations, according to Grubbs (1969) described in Newman (1995). This is good advice for environmental toxicologists.

(1) Investigators should reject any measurement they know has been obtained by a faulty procedure. The measurement should be rejected *whether or not* it appears to fit the presumed distribution. (The procedures responsible for all data should be scrutinized as part of the laboratory's normal program of quality control, whether or not those data appear unusual.)

(2) Next, investigators should consider that they might have adopted an unsuitable model, which might be the reason for poor fit of one or more observations. This possibility, often ignored, is an important one.

(3) Finally, if the anomaly remains unexplained, it should be reported, no matter what course is chosen for subsequent analysis of data.

10.2.1 Checking Errors and Procedures

Any apparent outliers should be re-checked for human errors. This includes measuring the effect, recording the data, transferring numbers, or entering into computer programs.

The happiest remedy for an aberrant datum would be to discover that it was caused by a mistake in transcription or arithmetic, which can be corrected immediately. *All other data-points should be checked in similar fashion*. There might also have been an error in a non-outlier, and a balanced scientific approach must subject all observations to the same scrutiny.

If no human slip of the pencil or keyboard were apparent, the investigator should look for biological or procedural items which could have caused the apparent anomaly. On the principle that "the test organism never lies", the investigator should consider all the potential environmental stimuli that the organisms encountered during their acclimation and testing.

The entire sequence of procedures should be scrutinized for correctness, in all parts of the test, and for all treatments in the test. This follows the first step listed previously.

10.2.2 Alternative Models

If no error is apparent, the next step could be to consider whether an unsuitable model was being assumed. For example, a smooth decrease in performance with increased concentration might be assumed, but the actual situation might be hormesis (increased performance at a low concentration, Section 10.3).

Another logical step in finding a more suitable model would be the possibility of transforming the data using a common procedure. A systematic trend in the data might be amenable to an advantageous transformation. For example, a general failure to show normality in the distribution might be remedied by the arcsine root transformation. If the outlier were a single aberrant point, the rationale for transformation would weaken, as would the likelihood of solving the difficulty by this method.

If transformation was of little assistance, analysis by nonparametric methods could be useful. A method that makes use of ranking can show good performance with an aberrant value, since it is usually less influenced by an outlier. The OECD (2004) suggests including such a nonparametric analysis (including the outlier) as a supplementary final step in a report which includes two parametric analyses (with and without the outlier, see following text).

A statistician might be able to provide a *robust model*, which uses a different *penalty function* (a rule for optimizing, such as minimum residual sums of squares) that minimizes the effect of the outlier. A comparison of the inferences obtained from regular and robust methods could guide decisions on the outlier.

10.2.3 Criteria for Outliers

In parallel with the procedures in subsections 10.2.1 and 10.2.2, an investigator should, if possible, use objective mathematical techniques to see whether the outlier appears to represent an anomaly or merely variation. Findings of these mathematical techniques must be tempered by judgement when they are

applied to toxicity results (see following text), but they can help to decide whether or not an anomalous value should be included in the overall analysis of results.

If there is no replication (as in the probit line for a quantal test), there are no objective means of identifying outliers. An investigator should use tabulated or graphed data in a report, to convey the magnitude of the anomaly.

With replicates, there are additional options. Figure 22 shows examples with possible outliers in the second-lowest concentrations. The value in the left panel of Figure 22 seems particularly divergent.

Rules of thumb. A rule of thumb might be applied for assessing an outlier among replicated measurements. If the observation is distant from the median by more than 1.5 times the inter-quartile range (i.q.r.), it is likely an outlier.[64] Unfortunately, the rule of thumb loses some usefulness in environmental toxicology because there are usually only a few replicates for any one treatment, and estimates of the inter-quartile range become rather shaky. For example, this procedure is not useful for the data in Figure 22, because it would be quixotic to estimate quartiles for a series of four measurements.

A variant of this informal procedure is Tukey's rule (Tukey, 1977), which includes somewhat fewer observations as possible outliers. A potential outlier would be 1.5 times the i.q.r. (or more) below the first quartile, or above the third quartile. From 1.5 to 3.0 times is considered a "mild outlier", and >3.0 times indicates a "severe outlier". This procedure has the same difficulty in deciding the inter-quartile range for the low numbers of replicates usually found in environmental toxicology. The OECD (2004) suggests that Tukey's rule might be used as a formal test by assessing the outliers in terms of the residuals (the results of subtracting the treatment mean from the individual values), in order to avoid confounding the outliers and treatment effects.

Statistical criteria for outliers. Statistical tests have been proposed for assessing a potential outlier in an objective manner. The method of Grubbs (1969) is recommended by Newman (1995). The suspected outlier in a group is entered into a formula with the mean of all observations (*x-bar*) and the standard deviation of all observations (SD), to estimate a value *T*. For environmental toxicology, "all observations" mean all values obtained at the particular concentration giving rise to the aberrant value. The formulae for high outliers and low outliers are as follows.

$$T = \frac{outlier - \bar{x}}{SD}$$

$$T = \frac{\bar{x} - outlier}{SD}$$

The value calculated for *T* is compared with a table of critical values provided by Grubbs (1969) and Newman (1995). If the calculated value exceeds the critical value, the suspected outlier is judged not to come from the same normal distribution as the rest of the values. [65] A key weakness of this parametric outlier test is the assumption that the data follow a specific distribution, in this case, normal. Whether or not a data-point is rejected depends on that assumption.

[64] Interquartile range is described in the Glossary and Appendix R. If five mean values in a series were 20, 24, 28, 34, and 40, the first quartile would be 24, the median 28, and the third quartile would be 34. The interquartile range would be 34 - 24 = 10. The criterion would be 10 times 1.5 = 15. The limits would be 28 ± 15 = 13 and 43. Both the lowest and highest values in the series are within the limits and probably not outliers.

[65] In his table, Grubbs (1969) gives a choice of three levels of significance, for up to 100 observations in the distribution. For environmental toxicology, lower numbers of observations would be common. For 3, 4, 5, ... 10 observations in the distribution, and a significance level of 5%, the critical values would be 1.15, 1.46, 1.67, 1.82, 1.94, 2.03, 2.11, and 2.18. The critical value for 20 observations would be 2.56, and that for 30 observations would be 2.75. These are one-tailed values, as would be appropriate for the formulae shown.

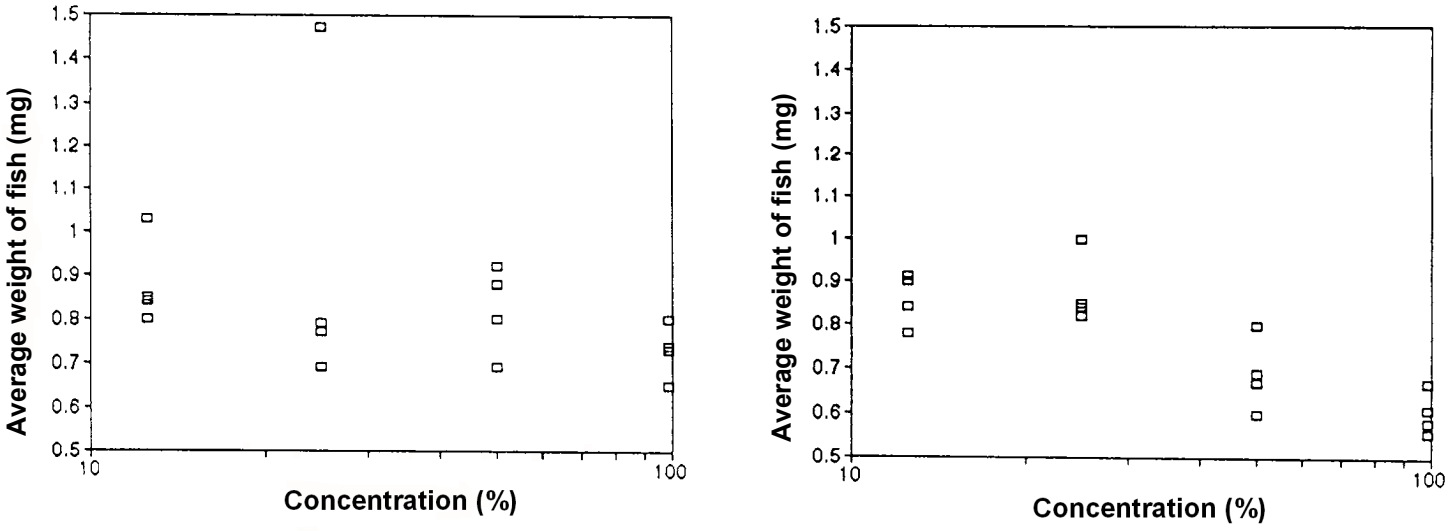


Figure 22 Examples of possible outliers in tests for seven-day growth of fathead minnow larvae. These data are from tests on two Canadian pulp mill effluents. In each example, one measurement at the second-lowest concentration (25% effluent) is above the other measurements at that concentration and also above the general distribution. There were four replicated at each concentration, each with nine or 10 larvae.

This formula can be applied to the data for the second-lowest concentration in the left panel of Figure 22, where there appears to be a definite outlier. The average weights of larval minnows from the four replicates are: 0.69, 0.77, 0.79, and a divergent 1.47 mg. Carrying extra significant digits for calculation, the mean is 0.93, standard deviation is 0.3626, and *T* is calculated as 1.49. The critical value (see footnote 65) for four measurements is 1.46. *T* is just greater than the critical value, so the high point could be classified as an outlier. This appears to be justified from inspecting the total set of data. It is noteworthy that *T* does not exceed the critical value as much as might be expected from the appearance of the graph.

Repeating this procedure for the right panel of Figure 22 shows that this objective test must be tempered with judgement when applied to data from toxicology. The four average weights of fish for the second-lowest concentration are 0.84, 0.82, 0.85, and 1.0 mg. The mean is 0.8775, standard deviation is 0.08261, and *T* is 1.48. Once again the critical value is 1.46. The calculated *T* just exceeds the critical value, so this provides some justification for rejecting the high value at this concentration, and proceeding with analysis using the other three. However, the entire distribution of data in the right panel of Figure 22 should be inspected. The overall variability in the second lowest concentration is not greatly different from variation in the other concentrations. The statistical decision on an outlier appears to have been driven by the tight clustering of three measurements, which are indeed unusually close together (0.84, 0.82, and 0.85). This clustering reduces the standard deviation to a very low value, and thus raises the calculated *T*. The assumed normal distribution might not be valid, a problem mentioned previously. Nor does this statistical method take into account the overall variation shown in the complete test; it was totally influenced by the minimal variation at the concentration of interest, which was apparently an unusual chance event. A method which incorporated total variation in a test would be superior for the needs of environmental toxicology.

The conservative approach in a marginal case like the right panel of Figure 22, would be to accept the questioned measurement. It is recommended that an investigator should provide analyses with and without the marginal point, along with a description of the situation and interpretive conclusions.

Other statistical methods for detecting outliers are described in the literature, but they seem to have the same weakness for use in toxicity tests, of not considering the total variation in all concentrations. The standard statistical text by Snedecor and Cochran (1980) gives two relatively straightforward formulae for testing whether an observation can be classified as an outlier. The methods use fairly detailed tables of the critical values in the tests, which cannot be reproduced here. There is also a dedicated book by Barnett and Lewis (1994), and the USEPA (1995) suggests consulting a publication by Draper and John (1981).

Multiple outliers. Application of a remedy becomes more doubtful for more than one suspected outlier at a single concentration. Collett (1991) advises that there "is no reliable objective procedure that can be recommended for assessing ... a group of two or more outliers"; however, a possible procedure is given by Rosner (1983). The same procedure is outlined by Newman (1995), along with code for a computer program in FORTRAN for the procedure. Snedecor and Cochran (1980) show how the two simple formulae for an outlier can, and should, be applied in the case of two outliers in a set of observations (say, in the replicates at one concentration). The most extreme of the outliers should be tested first. *Whether or not* the most extreme value was found to be a statistical outlier, that most extreme value should be removed from the set, and the second most extreme value should be tested within the remaining distribution of values. If it was an outlier, statistically, then *both* it and the most extreme value would be declared outliers. The rationale is that the most extreme value can "mask" the deviation of the second most extreme value, by its influence on the overall distribution.

Although there is no completely adequate criterion for outliers in toxicity tests, the method of Grubbs (1969), outlined previously, seems as suitable as any. Statistical methods should be developed that are more appropriate for toxicology. Meanwhile, if an apparent outlier is critical to interpreting a test one way or the other, the advice of a statistician should be sought to apply measures, not described here, quantifying the degree of influence that a particular observation has for a model.

10.2.4 Actions for Reporting

If an outlier is suspected, the sequence of desirable approaches and actions can be summarized.

- All records of the test should be examined for errors in observation or recording.
- Next, the procedures used in holding and testing should be reviewed to see if they had triggered some understandable biological response.
- If not, alternative models for the results should be considered.

In parallel with these steps, the investigator should use objective statistical methods to examine the question of rejecting or accepting the variant observation(s).

In reporting, anomalies should be listed and results and conclusions of the investigative steps should be described.

If statistical techniques do *not* indicate an outlier, that should be reported, and the overall analysis of results should proceed with the aberrant result included. (An analysis without the anomalous value might also be included in the report, with comments on any influence on interpretation.)

It should also be reported if the statistical test identifies an outlier. Results should be analyzed with and without the questionable value. Both analyses should be reported, the investigator should indicate which one is chosen as definitive, and should offer a commentary on reasons for the choice. An alternative or additional analysis by a nonparametric or other more robust method could provide additional enlightenment. This duplication of analysis and explanation might not fit some regulatory programs, which usually require one standardized result. In that case, investigators should report the best estimate in their judgement, with an indication that additional background analyses and explanations are attached or held on file.

10.3 Hormesis—Stimulation at Low Concentrations

Key Guidance

- *There are many sublethal tests in which performance is stimulated at low concentrations, i.e., hormesis or "better than the control". This brings a philosophical problem of deciding whether these are deleterious effects, and which level of performance should be considered the control. No general answer prevails.*
- *Hormesis also brings practical problems of analysis. Standard dose-effect models do not fit, or else they cause fallacious estimates. More complex models can lose power for detecting the deleterious concentrations, if a minimal experimental design is adopted.*
- *For point estimates such as IC25, the best approach is to fit the data with a nonlinear regression, and estimate the IC25 by comparison to the true control. A standard analytical approach for this is included in new methods from Environment Canada and in this document.*
- *If hypothesis testing is done, normal procedure should be followed, with data included from all concentrations. However, only effects significantly "worse" than the true control should be considered in designating the NOEC/LOEC.*
- *For tests showing hormesis, reports should include the original results and explanation of methods of analysis.*

In hormesis, a low concentration of the test material acts as a stimulant for performance of the test organisms, compared to the control organisms, i.e., they perform "better" than the control. At higher concentrations, deleterious effects are seen. The more general category of "low-dose stimulation" is usually the more appropriate term to use. It includes other possible causes of stimulation, such as solvent effects, experimental error, or conceivably, just a general arousal of test organisms held under monotonous laboratory conditions ("sufficient challenge"). Low-dose stimulation is sometimes seen in a variety of effects including increased growth of test organisms, or increased algal cell density, shown in Figure 23.

Low-dose stimulation is, perhaps, the most common hindrance to analysis of well-designed sublethal toxicity tests. It represents a real phenomenon, not outliers or a flawed test, and some Canadian investigators encounter waste materials that reliably produce hormetic results. This is not limited to environmental studies; it is widespread in medical toxicology (Davis and Svendsgaard, 1990). Calabrese and Baldwin (1997) reviewed positive effects ranging from 30 to 60%, although Canadian environmental laboratories experience +30% as a more likely maximum.

The cause of increased performance is seldom determined. For growth or number of cells produced, the findings would be consistent with some nutrient being added with the test material, stimulating production. If that were so, and if the nutrient were known, the obvious remedy would be to add the nutrient itself to all concentrations including the control. The level of understanding would rarely justify such a remedy.

10.3.1 The Difficulties

Problems with usual methods. If low-dose stimulation is present, and common techniques of analysis are applied, the endpoint usually tends to be lowered ("more toxic"). The control or baseline performance is often over-estimated, leading to an over-estimation of the effects of concentrations and a lowering of ICp. The slope of the fitted relationship usually becomes steeper, which could affect the estimate of confidence limits.

If the data shown in Figure 23 are entered without change into the ICPIN program (Section 6.4.3), the standard smoothing process raises the control to a higher number of cells and estimates a lower ICp than might be expected. The original (actual) control value of 2650 cells is adjusted to 2860 cells. Smoothing, assigns the same value of 2860 cells to each of the first four concentrations. Accordingly, if the IC25 is being estimated, it is based on about 2145 cells (75% of 2860) instead of the original 1988. The IC25 is estimated as approximately logarithm 3.92 or 8300 concentration units, compared to logarithm 4.05 or 11 220

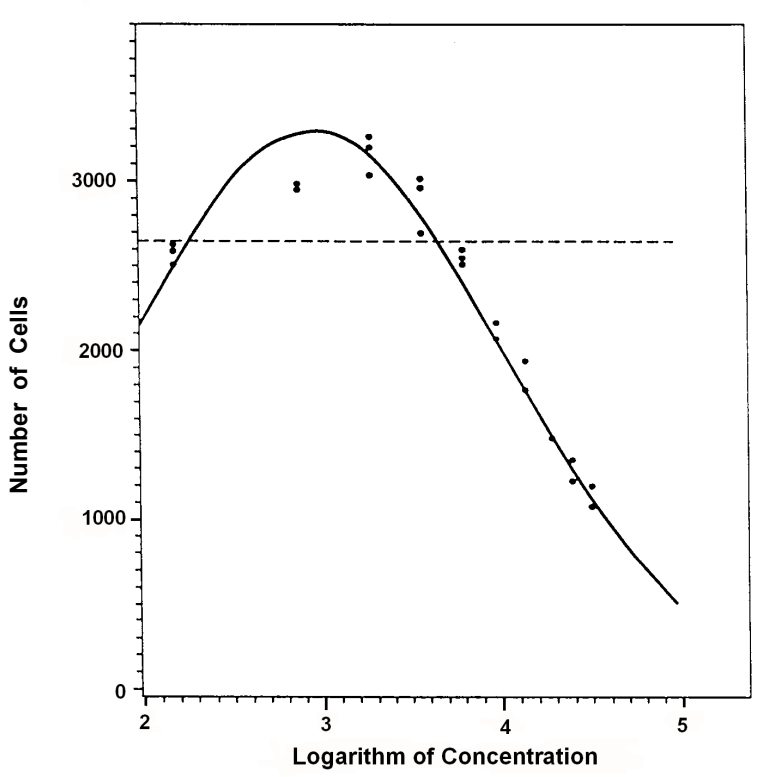


Figure 23 An example of stimulation at low concentration. This was a test with the green alga, *Pseudokirchneriella subcapitata [*formerly *Selenastrum capricornutum]* in a Canadian laboratory. The horizontal dotted line shows performance of the control, and the lowest concentration was similar. The alga showed increased reproduction at the next three concentrations, then the expected decrease at high concentrations.

concentration units if the original control had been used, which is certainly an appreciable change.

From this, it is clearly not desirable to simply ignore low-dose stimulation and proceed with routine use of commonly used statistical methods.

Effect or enhanced control? There is a philosophical conundrum in dealing with hormesis, and there is no consensual solution. There are several possible approaches, but none of them are completely satisfactory.

Should the improved performance be considered an "effect" of the toxicant and therefore, by definition, undesirable? Outside the laboratory, in a living community, diversion of an organism's energy into avenues such as growth might indeed be deleterious.

It might steal energy that would be more strategically used for reproduction or some other activity. However, within a laboratory test, such speculation is hard to support; it is difficult to consider improved performance in the test criterion as a *harmful* effect.

On the other hand, if low-dose stimulation were adopted as a deleterious effect for the example in Figure 23, it would mean that the 2nd, 3rd, and 4th concentrations would be declared as potential toxic effects, which would probably not seem rational to most observers.

Still another alternative would be to regard the improved performance as a sort of enhanced control, perhaps resulting from improved supply of nutrients. Or perhaps for animals, it might be a response to "sufficient stimulus", compared to an otherwise monotonous set of conditions for existence under the control conditions. The stimulated measurement would then represent potential performance, and assume the function of the control. Most investigators would probably consider that unrealistic, and instead, opt for comparison with the standard control.

It is not desirable to adopt some blend of control and stimulated performance as a new control, as was done in the smoothing of the ICPIN program. This method, as previously shown, results in an appreciably lowered endpoint.

Figure 23 shows the difficulty of answering such fundamental questions, and in the past there has not been a real consensus on these matters.

Aside from the philosophical questions, low-dose stimulation brings very practical problems of deriving a statistical approach to analyze the data (see following options). One potential problem is that hormetic effects at low concentrations would mean that the statistical procedure had fewer concentrations showing decreased performance for modelling the reduction from the control.

10.3.2 Including Hormetic Effects in Regression

Environment Canada has adopted this reasonable option in its recent standard methods of nonlinear regression to estimate ICps (EC, 2004a–c; Section 6.5.8 and Appendix O). The true control is used for comparisons of effect. A special model accommodating hormesis is used, and the procedure resolves the analytical problems referred to previously. This felicitous solution to the statistical part of the hormesis problem has many advantages. There is no need to reject any of the data. There is no need to change the control value by smoothing or other remedies. There is no distortion of effects at concentrations higher than those showing hormesis. This is listed in Section 10.3.3 as "Option 1" for analyzing data which show low-dose stimulation.

Generalized linear models (GLIMs) have been applied to the results of *Ceriodaphnia* tests, and showed promise for dealing with low-dose stimulation which often occurred (Bailer *et al.,* 2000a). Using GLIMs allowed more consistent estimates of ICp than with the program ICPIN. The GLIMs were equally applicable to data that were quantal, quantitative, or counts.

Brain and Cousens (1989) described some hormetic sets of results by reparameterized logistic (sigmoid) models. The added parameter of the equation allowed for a hormetic change in performance at low concentrations. The approach was further developed by including the desired endpoint as a parameter (van Ewijk and Hoekstra, 1993), and that technique has been incorporated into the recent methods of Environment Canada. The advantage is direct estimation of the endpoint and confidence limits from the data. A possible disadvantage is the need to estimate four parameters using nonlinear regression, requiring an experimental design that produces a data-set with adequate numbers of concentrations and replicates. The parameterization of van Ewijk and Hoekstra (1993) is sensitive to the optimization algorithm underlying the nonlinear package (B.A. Zajdlik, 2004, pers. comm. B. Zajdlik & Associates Inc., Rockwood, Ont.).

The more sophisticated method of analysis eliminates the statistical part of the hormesis conundrum. It solves the philosophical problems of Section 10.3.1 by the reasonable approach of using the performance in the true control as the basis for judging effects (Option 1, Section 10.3.3). The philosophical question of what should be designated as "normal" performance might still be debated in some unusual situations, and that is considered in Section 10.3.3.

10.3.3 Options for Dealing with Hormesis

A range of approaches is outlined in the following options for tests with stimulation at one or two low concentrations. Option (1) is recommended, and is the required first choice in Environment Canada's recent soil tests (EC, 2004a–c). Option (4) is recommended if it is necessary to make point-estimates by the ICPIN program, and Option (5) for estimating NOEC/LOEC, if hypothesis testing is used for some reason.

(Option 1) In point estimates, include hormesis in a more complex model. If an IC25 is being estimated, adopt the model for hormesis and carry out nonlinear regression (Section 6.5.8). The IC25 is still estimated in relation to the true control performance.

(Option 2) Smooth the effects for the control and low concentrations. Smoothing is done in the commonly used computer program ICPIN, to estimate the ICp. This adjusts the control to "better" levels, with a consequent lowering of the estimated ICp. A similar result could occur in hypothesis testing. This option is not recommended, because it makes comparisons with a control that does not, in fact, exist.

(Option 3) Omit concentrations showing significant hormesis from the statistical analysis. There is no mathematical or statistical basis for using this option; it could only be considered biological

judgement. This option might fit estimation of ICp by the ICPIN method. It would not be suitable for point estimates by regression. This technique would require a preliminary analysis to decide which concentrations actually demonstrated hormesis. Once the data-points were removed, analysis could proceed on the remaining concentrations. Accompanying the analysis, there would have to be a clear statement of the values omitted, and why.

This option is followed in only one of Environment Canada's toxicity test methods, the method for inhibition of algal growth (EC, 1992d). If algal growth in a test concentration is greater than in the control, those observations are reported but are not used for calculating the ICp.

This option would be unsatisfactory under certain circumstances. It could estimate an endpoint that was unrealistically low, if removal of the hormetic data left a wide "gap" between two low concentrations that spanned the endpoint. A hypothetical example is shown in footnote 66 in line no. 2 of the table [66].

[66] This hypothetical example represents the number of algal cells counted at various test concentrations. For simplicity, no replicates are shown.

	Control	6 (mg/L)	12 (mg/L)	25 (mg/L)	50 (mg/L)	100 (mg/L)
1. Observed number of cells	200	200	275	300	100	50
2. Hormetic concentrations removed	200	200			100	50
3. Control value substituted for hormesis	200	200	200	200	100	50

The observed effect in the first line indicates clear hormetic effects at 12 mg/L and 25 mg/L. These two effects are omitted in the second line, as in Option (3) of the text. If one intended to estimate the IC25 (concentration resulting in 150 cells) using the ICPIN method and the data in line 2, there would be interpolation between 6 mg/L and 50 mg/L, values which are rather widely separated. The ICp would be estimated as 17 mg/L, unrealistically low since there was no evidence of damage to algal production in the original data for 25 mg/L.

The third line of the table shows a procedure that has apparently been used by some laboratories to force a more realistic endpoint (Option 4 of the text). The control value is arbitrarily assigned to the concentrations that showed hormesis. The interpolation of the endpoint would now be between 25 mg/L and 50 mg/L. The IC25 would be 35 mg/L, which appears to be more reasonable.

Because of this uncertain usefulness, the method is not recommended as a complete solution for tests other than algal growth.

(Option 4) Assign the control value to concentrations showing low-dose stimulation. This is arbitrary and has no statistical justification, but is likely to provide realistic endpoints, with straightforward calculations by commonly used methods. It is not recommended for point-estimates, because a suitable method for regression is available (Section 6.5.8). Option (4) would function for point-estimates with the ICPIN program. This option was once supported for transitory use by a cross-section of Canadian investigators (Schroeder and Scroggins, 2001), but only until suitable regression methods were developed, as they have been now.

(Option 5) In hypothesis testing, consider low-dose stimulation as non-harmful. The statistical analysis would proceed as usual, i.e., it would include performance higher than the control. If the analysis showed that one or more low concentrations showed significantly better performance than the control, that information would be reported but it would not be considered a deleterious effect. The LOEC would be designated as the lowest concentration which resulted in a significant *decrease* in performance compared to the control. The endpoint would be the same as would be obtained with Option (4), but is preferred for hypothesis testing since it does not involve any manipulation of the original data.

These options might not be appropriate for all results. Investigators should inspect the graphed data to judge reasonable approaches and endpoints.

In all cases of low-dose stimulation it is important to:

- report the original data, and
- state which measures were adopted for analysis.

10.4 Deviant Concentration-effect Relationships

Key Guidance

- *Examples of several unusual or difficult types of data-sets are given. Suggestions are offered on interpretation.*

- *Some of the difficulties can be avoided by appropriate test design, particularly using a wide enough spread of concentrations.*

Most laboratories encounter unusual concentration-effect relationships occasionally. Test organisms seldom lie, so aberrant results usually have an explanation, but it might not be obvious. This section shows graphs of some unusual findings, with possible explanations and recommendations for dealing with the results. Initial interpretation should use plotted graphs, as recommended in Sections 4.2.2 and 6.3.1 and in guidance from the USEPA (2000a).

The series starts with some "good" data, for comparison. Some anomalous examples were obtained in Canadian laboratories during regular test programs, others are patterned after examples used by the USEPA (2000a). Problems of outliers and hormesis are covered in preceding sections.

(1) Good concentration-effect data. A regular linear relationship for an algal test is shown in Figure 24. There is no difficulty in estimating an endpoint such as an IC25 by various linear approaches. Hypothesis testing also performs satisfactorily.

The good results shown in Figure 24 would be gratifying to obtain, but they could have been improved at the design stage. There were eight concentrations tested within an order of magnitude, i.e., the concentrations were spaced close together. As mentioned in Section 2.2, such a design runs a risk of missing concentrations of interest, and in fact, that happened in this example. The lowest tested concentration shows a result about 13% lower than the result for the control. One facet of good design is to have at least one low concentration yielding results that are essentially the same as those of the control. The concentrations should have been spread over a wider range.

The results in Figure 24 are generally monotonic, and the slightly irregular effect at the third highest concentration would not cause concern. Presumably it represents background variability and would contribute a slightly larger variance to any statistical description of a fitted line.

A similar and remarkably straight concentration-effect relationship is shown in Figure 25. There would be no difficulty in analyzing such results, either for a point-estimate or hypothesis testing. It is somewhat unusual that the effects from low to high are spread out over more than two orders of magnitude. The spread, however, would not impede the analysis; the design was adequate, and both low and high effects were obtained. An additional lower concentration in the series, however, might have shown an effect that was closer to the control effect. Again, this good result illustrates the importance of designing tests with a wide spread of concentrations, rather than trying to guess which narrow range will be the important one (Section 2.2). In this test, the surprisingly wide spread of effects encompassed all of the concentrations of a design that would normally be considered adequate in breadth.

(2) Steep relationships. Sharp changes in effect at successive concentrations are common in environmental toxicity tests. The example in Figure 26 is not quite "all-or-nothing", because there is one intermediate effect as the relationship changes from a control value to a major deleterious effect. This type of data is moderately satisfactory; and the estimated endpoint will be reasonably precise, with narrow confidence limits (depending on the dilution factor used to choose concentrations).

A desirable feature shown in Figure 26 is the presence of a low concentration with an effect similar to the control value. This is one indication of suitable design and procedure in the test. There are, in fact, four low concentrations similar to the control, and statisticians would point to improved precision if more of the data-points had been in the region of rapid change. Accordingly, an improved design, in this case, would have omitted some of the lowest concentrations, in order to provide more data at the higher concentrations. Ideally, a range-finding test would have indicated the appropriate series of concentrations for the definitive toxicity test. However, lacking a range-finder, any change in design of the test concentrations would represent hindsight. As pointed out previously, a design that narrows the concentration range can be dangerous in testing a material of unknown toxicity. Some important concentrations might be "missed", and so it is better to spread the concentrations as was done in this test.

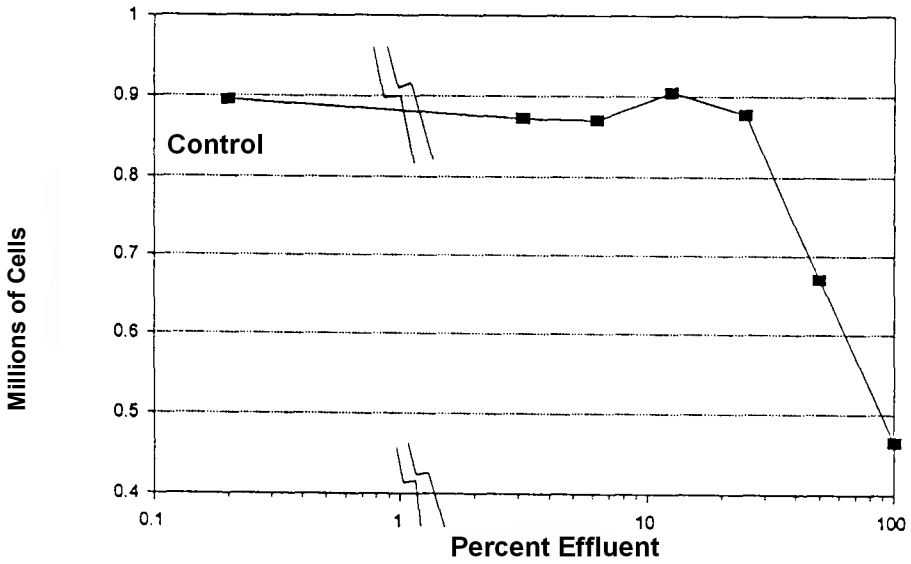


Figure 24 **An example of a good linear relationship of concentration and effect.** These are results for a Canadian test of toxic surface water with the alga *Pseudokirchneriella subcapitata [*formerly *Selenastrum capricornutum]*. For convenience of illustration on the logarithmic scale, the control is plotted as a very low concentration. Zigzag lines represent the discontinuity in use of the concentration scale.

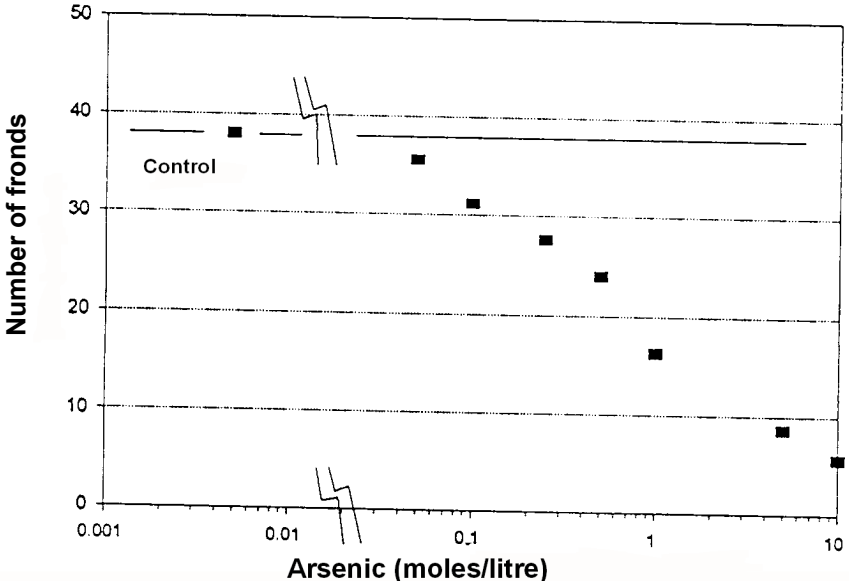


Figure 25 **Another example of a good relationship of concentration and effect.** Results from a Canadian laboratory, for growth of duckweed *(Lemna minor)* in concentrations of arsenic. (Other description as in Figure 24.)

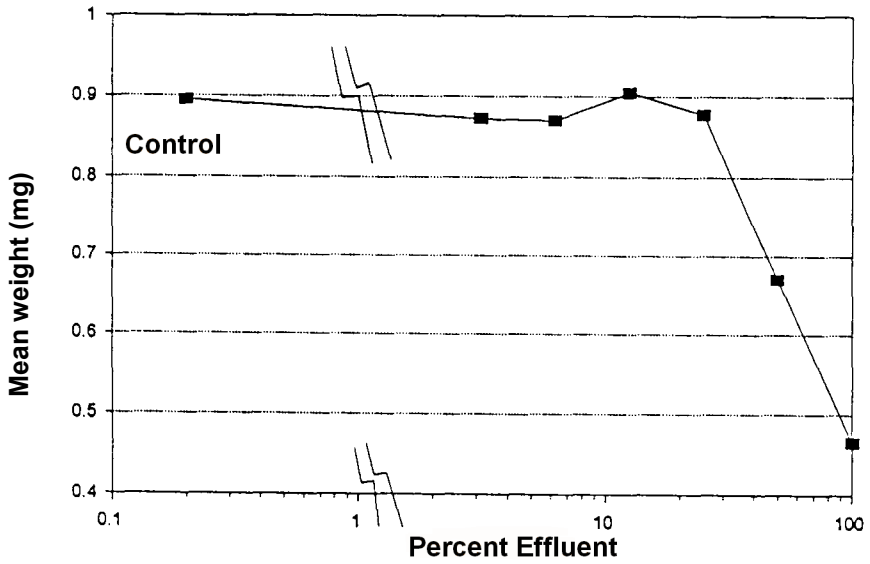


Figure 26 A steep relationship for weight of fathead minnow larvae exposed to concentrations of an effluent. Results from a Canadian laboratory. (Other description as in Figure 24.)

(3) Lack of effect with an irregularity. Sometimes no effect of the test material is evident at the highest concentration tested. If the material is an effluent or a sample of sediment or soil, nothing higher than 100% concentration can be tested. Interpretation is simple -- no harmful effect was demonstrated in this test. No point-estimate (ICp) can be calculated, and hypothesis testing would also show no effect.

The results in Figure 27 illustrate this, but show one inconsistency. The middle concentration is appreciably lower than the control level. Rarely, a laboratory might encounter such a pattern in a sublethal test. At the anomalous concentration, performance might be 25% worse than the control, and might also be statistically different from the control.

If analysis of the tests is normally done by point estimates, this irregularity does not cause a problem. The low value would not result in an endpoint and should be reported as an anomaly. If an investigator intended to use hypothesis testing, the irregular effect at the middle concentration might emerge as the LOEC. The lack of effect at higher concentrations invalidates any such estimate. The only reasonable approach would be to acknowledge the apparent anomaly, and state that hypothesis testing was not appropriate.

An explanation should be sought. Examine the records for divergent test conditions such as pH or dissolved oxygen. A single divergent replicate might have influenced the mean value (see outliers, Section 10.2). It is possible that failure to randomize might have influenced results through condition of organisms or some factor related to position in the array. If no explanation can be found, there is little option except to describe the range of results that were obtained, with a conclusion of one anomalous data-point.

(4) Anomalous intermediate lack of effect. Sometimes, an apparent progressive increase in effect is interrupted by a concentration showing lack of effect, similar to the control (Figure 28). Analysis could be carried out by methods which estimate an ICp. Line-fitting techniques would take the irregularities into account and produce appropriately wide confidence limits. ICPIN would force monotonicity on the relationship (Norberg-King, 1993), probably with satisfactory analysis in this case. Hypothesis testing would be spoiled if

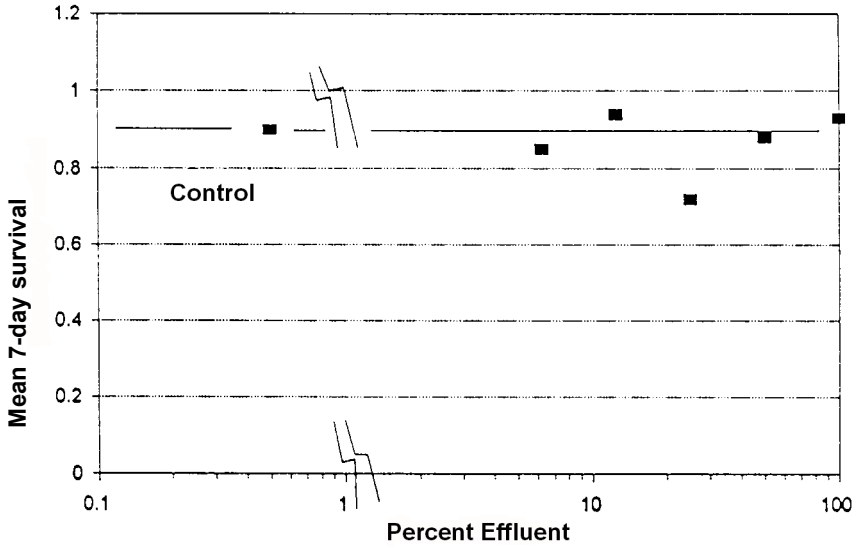


Figure 27 Lack of effect at high concentrations, with an anomalous intermediate concentration. Hypothetical data for survival of fathead minnow larvae.

the anomalous point were significantly different from the control; two sets of NOEC/LOEC would be generated. In that case, USEPA (2000a) recommends choosing the lower one as the NOEC (6.25% in Figure 28) if the test shows a satisfactory MSD (see Section 7.2.4). That would be a satisfactorily cautious approach.

The anomalous pattern should be reported, whether or not there is a successful point estimate. A search of test procedures should be made for a cause, as in example (3).

Rarely, a biological reason might be ascertained for anomalous dose-effect situations, and aggression is an example. In a series of 90 lethal screening tests on an industrial effluent, a few tests were seen to have extreme aggression among trout, after they were put into the test tanks. Particularly odd results were obtained in two tests. Out of the totals of 20 fish per treatment, 9 died in the control, and 5 in the lowest concentration, apparently the result of fighting. There were no deaths in two intermediate concentrations, in which the fish appeared to be pacified by the effluent. In full-strength effluent, toxicity came into play and 16 fish died (Sprague, 1995). The major effect in the control was a clear message that there was some extraneous factor acting in the test. The deviant U-shaped relationship of mortality to concentration could be explained but not analyzed by conventional means.

(5) Flat effect-curve. Figure 29 indicates an apparent slight effect at many concentrations, but no increase in effect with higher concentration. Clearly, there is some anomaly. Results on the right-hand side might or might not be significantly lower than the control, but the consistent pattern should alert the investigator. There should be no attempt to estimate an endpoint for data as extreme as those illustrated.

A search should be made for a procedural or biological reason. Some possibilities follow.

(a) The control performance might be unusually high. It should be compared with previous control findings at the laboratory. If high, the test results merely indicate that the tested effluent is not toxic at any concentration. (It is unlikely that this would represent a control performance that was unusually low. If that were the case, it would mean that most or all of the test concentrations caused an effect, but without a concentration-effect relationship.)

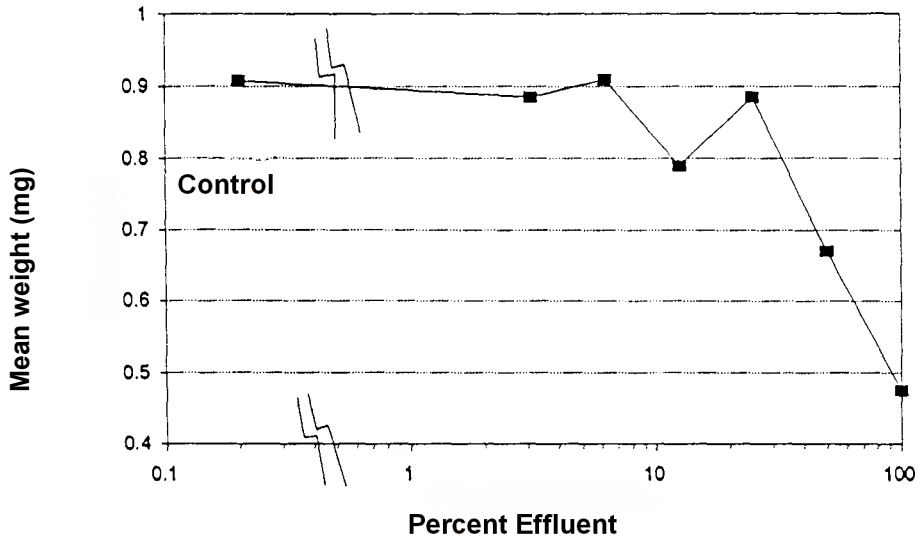


Figure 28 **An apparent anomalous lack of effect at an intermediate concentration.** Hypothetical example, modified from data in Figure 26 for weight of larval fathead minnows.

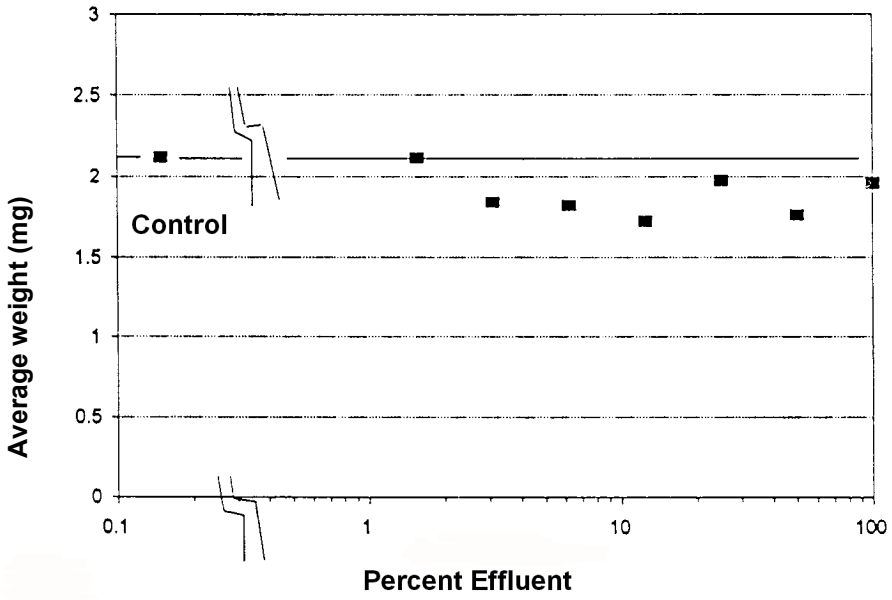


Figure 29 **An apparent slight effect, but flat with concentrations.** This example is from a Canadian laboratory using the test to measure larval weight of silversides, a marine fish.

(b) Dilution waters might be inappropriate. If the control used one type of water (say the culture water) and the test concentrations used some other water, that could be a logical explanation of a flat, lowered distribution. Such a situation should not arise since procedures of Environment Canada call for a single control/dilution water. In the example of Figure 29, the explanation might lie in some effect of the seawater brine or salts used to adjust salinities in the test concentrations.

(c) There might be pathogenic effects. This is unlikely but might occur in chronic tests, especially with fish. There might be pathogens in the tested material, which caused a low-grade effect on the organisms, although the material itself was not toxic. If this happened, results would probably be more erratic than shown in Figure 27. If pathogens seemed likely, and it were desired to investigate, parallel tests might be run with one set using ultraviolet or antibiotic treatment on the test material.

If this pattern persisted in a testing program, it might be desirable to investigate by chemical analysis or toxicity identification techniques.

(6) Inverse relationship of effect to concentration. At first glance, Figure 30 might seem like a suitable regular relationship. A second glance reveals that performance of the algal cultures improves as concentration increases. The conclusion is straightforward; the effluent being tested is not toxic to algae, but is providing some nutrient which enhances their growth and reproduction. Such an effect-curve is most likely to be seen with plants, but might conceivably occur with other organisms. (The evidence of nutrients in the tested material should be considered from a wider perspective, with regard to enrichment of the receiving waterbodies.)

Another unlikely but possible explanation would be that the tested material was not toxic, but the control/dilution water was. If the receiving water was used for dilution, it would appear to be toxic already. If this were a possible explanation, and the "absolute" toxicity of the effluent or other test material is to be determined, then a standard dilution water should be used, one known to be favourable for the organisms.

(7) Strong effects at all concentrations. The example in Figure 31 shows major effects on algal numbers at all of the tested concentrations. There is also a very flat relationship of effect to concentration. This is a real test result, not a hypothetical situation. Obviously a wider range of concentrations should have been tested, as discussed under items (1) and (2). The five tested concentrations spanned only one order of magnitude. If they had been spread more widely, the pattern of results might have been less enigmatic. A dependable IC25 cannot be estimated with the present data, nor would there be a reliable approach for NOEC/LOEC.

It is difficult to judge whether the effects might have extended into higher and lower concentrations because of the limited range of concentrations shown for Figure 31. To explain the flat distribution, one might stretch for an explanation involving some balance within the material being tested, between toxic components and others that stimulated algal growth. Perhaps there could be a chemical explanation in the amount of active form or component that was free to operate at the various concentrations.

10.5 Procedural Interactions that Affect Results

Key Guidance

- *Procedures selected for testing could influence the analysis and results.*
- *In a growth test, for example, some deaths in a container at high concentration could result in the remaining organisms having a bigger share of food, showing compensatory growth, and thus compromising the analysis and obscuring sublethal effects.*
- *Similarly, partial mortality in a container could result in more*

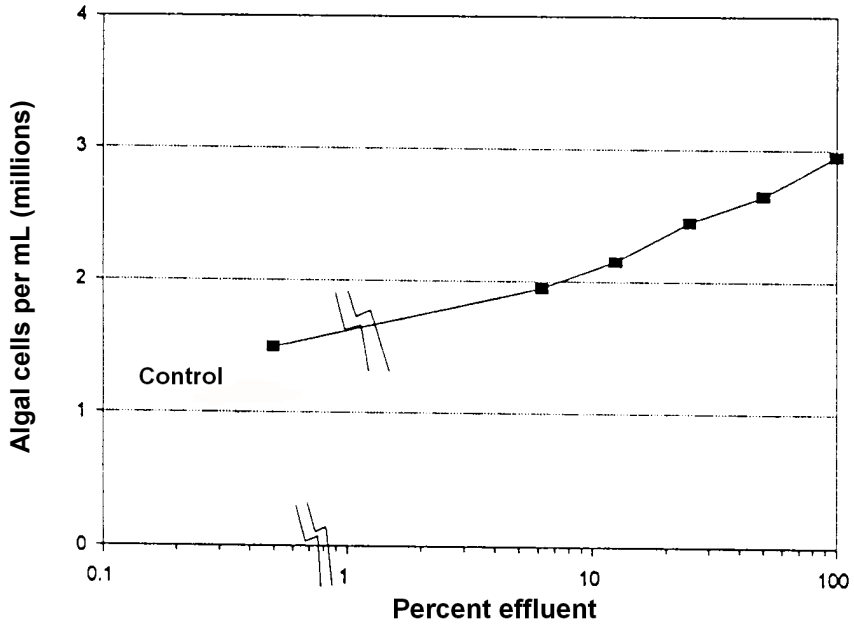


Figure 30 **An example of better performance with higher concentration.** The example is for number of algal cells and is taken from USEPA (2000a).

toxicant available to the survivors, enhancing sublethal actions.

- *The best defence against such influences is to use proven procedures. Selection of feeding regime and favourable renewal rates for test solution might avoid the problems mentioned.*

Questions may arise concerning the influence of the test methods themselves on the statistical analysis and estimate of endpoint. The interaction can produce results that are inappropriate for statistical analysis or difficult to interpret. This topic is seldom considered, but could sometimes be important. One example is given here; there would be parallel examples for other tests and other effects.

The number of organisms per container could easily affect analysis and results in sublethal tests. If there were several test organisms per container, and they were dying in some containers and not in others, that could lead to unbalanced exposures influencing effects. The unequal treatment of groups could violate requirements for statistical analysis.

- If most of the organisms died in one container, would the remaining ones benefit from more available food? Could an assessment of growth be biased? It is certainly possible, if the feeding technique provided more food per organism in those containers with deaths. There are clear examples of compensatory (increased) food intake in fish, which can overcome the deleterious effect of a toxicant on growth (Warren, 1971). Use of a feeding rate based on number or biomass of organisms could rectify this problem.

- The choice of feeding rate could easily influence the result obtained, in view of the previously noted phenomenon of compensatory growth. An investigator might choose a relatively high feeding rate in hopes of showing maximum differences among the test concentrations, but too high a rate could smooth differences because of the compensatory feeding.

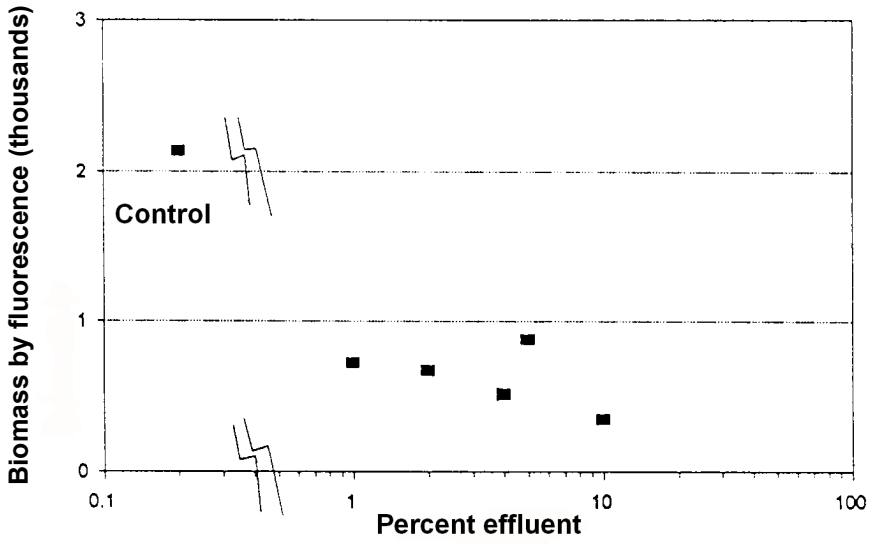


Figure 31 **Test results showing only strong effects.** The example from a Canadian laboratory is for biomass of the alga *Pseudokirchneriella subcapitata [* formerly *Selenastrum capricornutum]*.

- If most of the organisms in a container died, would the remaining ones receive greater exposure (dose)? Presumably, there would no longer be uptake of toxicant by the organisms that died, so they would not be lowering the ambient concentrations. The exposure of the remaining organisms would be greater than a situation of no mortalities. The effect might or might not be negligible.

The main defence against such anomalous results is to use good, well-tested test procedures. Standard methods such as those of Environment Canada are widely available now; the methods are beneficial, having generally been hammered out by experienced groups of investigators. Feeding rates would have been chosen to minimize anomalies. For maintaining desired concentrations in the test containers, any influence from partial mortalities in groups of organisms would be overcome by favourable high volumes of test solution for the biomass involved.

References

Abbott, W.S., 1925. A method of computing the effectiveness of an insecticide. *J. Econ. Ent., 18:*265–267.

Alderdice, D.F. and J.R. Brett, 1957. Some effects of kraft mill effluent on young Pacific salmon. *J. Fish. Res. Board Can., 14:*783–795.

Andersen, H. 1994. Statistical methods for evaluation of the toxicity of waste water. M.Sc. thesis, Dept. of Mathematical Modelling, Technical Univ. of Denmark, Lyngby, Denmark. [In Danish]

Andersen, J.S., H. Holst, H. Spliid, H. Andersen, A. Baun, and N. Nyholm, 1998. Continuous ecotoxicological data evaluated relative to a control response. *J. Agric. Biol. and Environ. Statistics, 3:*405–420.

Andersen, J.S., J.J.M. Bedaux, S.A.L.M. Kooijman, and H. Holst, 2000. The influence of design characteristics on statistical inference in non-linear estimation; a simulation study. *J. Agric. Biol. and Environ. Statistics, 5:*28–48.

Anon., 1994. How to measure no-effect? SETAC News, Nov. 1994: p 19. [Society Environ. Toxicol. and Chemistry]

APHA, AWWA, and WEF, 1992 [Amer. Public Health Assoc., Amer. Water Works Assoc., and Water Environ. Fed.]. *Standard methods for the examination of water and wastewater.* 18th ed. APHA, Washington, D.C.

Ashton, W.D., 1972. *The logit transformation with special reference to its uses in bioassay.* Griffin's Statistical Monographs & Courses, no. 332. Hafner Pub. Co., New York, N.Y., 88 p.

Atkinson, G.F., 1999. Assessment of available computer programs. Attachment T, 2 p., In: *Minutes/Proceedings of the Statistics Workshop for Toxicological Testing,* Pacific Environmental Science Centre (PESC), North Vancouver B.C., September 15–17th, 1999. Environment Canada, Pacific Environmental Science Centre, North Vancouver, B.C.

Bailer, A.J and J.T. Oris, 1993. Modeling reproductive toxicity in *Ceriodaphnia* tests. *Environ. Toxicol. Chem. 12:*787–791.

————. 1994. Assessing toxicity of pollutants in aquatic systems. p. 28–40, In: *Case studies in biometry.* N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse (eds). John Wiley & Sons, Inc., New York, N.Y.

———. 1997. Estimating inhibition concentrations for different response scales using generalized linear models. *Environ. Toxicol. Chem., 16:*1554–1559.

———. 1999. What is an NOEC? Non-monotonic concentration-response patterns want to know. *SETAC News, March 1999:*22–24.

Bailer, A.J., M.R. Hughes, D.L. Denton, and J.T. Oris, 2000a. An empirical comparison of effective concentration estimators for evaluating aquatic toxicity test responses. *Environ. Toxicol. Chem., 19:*141–150.

Bailer, A.J., R.T. Elmore, B.J. Shumate, and J.T. Oris, 2000b. Simulation study of characteristics of statistical estimators of inhibition concentration. *Environ. Toxicol. Chem., 19:*3068–3073.

Baird, R.B., R. Berger, and J. Gully, 1995. Improvements in point estimation methods and application to controlling aquatic toxicity test reliability. p. 103–130, In: *Whole effluent toxicity testing as evaluation of methods and prediction of receiving system impacts.* D.R. Grothe, K.L. Dickson, and K.K. Reed-Judkins (eds.), SETAC Press, Pensacola, Fla.

Barnett, V. and F. Lewis, 1994. *Outliers in statistical data.* 3rd ed. Wiley, New York, N.Y.

Bartlett, M.S., 1937. Some examples of statistical methods of research in agriculture and applied biology. *J. Roy. Stat. Soc. Suppl., 4:*137–170.

Bates, D.M. and D.G. Watts, 1988. *Nonlinear regression analysis and its applications.* John Wiley & Sons, New York, N.Y. 365 p.

Beyers, D.W., T.J. Keefe, and C.A. Carlson, 1994. Toxicity of Carbaryl and Malathion to two federally endangered fishes, as estimated by regression and ANOVA. *Environ. Toxicol. Chem., 13:*101–107.

Billington, J.W., G.-L Huang, F. Szeto, W.Y. Shiu, and D. MacKay, 1988. Preparation of aqueous solutions of sparingly soluble organic substances: I. Single component systems. *Environ. Toxicol. Chem., 7:*117–124.

Bliss, C.I., 1937. The calculation of the time-mortality curve. *Ann. Appl. Biol., 24:*815–852.

Bliss, C.I. and McK. Cattell, 1943. Biological assay. *Ann. Rev. Physiol., 5:*479–539.

Borgmann, U., 1994. Chronic toxicity of ammonia to the amphipod *Hyalella azteca*; importance of ammonium ion and water hardness. *Environ. Pollut., 86:*329–335.

Brain, P. and R. Cousens, 1989. An equation to describe dose responses where there is stimulation of growth at low doses. *Weed Res., 29:*93–96.

Broderius, S.J., 1991. Modeling the joint toxicity of xenobiotics to aquatic organisms: basic concepts and approaches. p. 107–127, In: *Aquatic toxicology and risk assessment: fourteenth volume.* ASTM STP 1124, M.A. Mayes and M.G. Barron (eds.), Amer. Soc. Testing and Materials, Philadelphia, Pa.

Bruce, R.D. and D.J. Versteeg, 1992. A statistical procedure for modeling continuous toxicity data. *Environ. Toxicol. Chem., 11:*1485–1494.

Buikema, A.L., Jr., B.R. Niederlehner, and J. Cairns, Jr., 1982. Biological monitoring. Part IV -- Toxicity testing. *Water Res., 16:*239–262.

Burchfield, R.W., 1996. *The new Fowler's modern English usage*. 3rd ed. Clarendon Press, Oxford.

Calabrese, E.J. and L.A. Baldwin, 1997. The dose determines the stimulation (and poison). Development of a chemical hormesis database. *Int. J. Toxicol., 16:*545–559.

Calamari, D., R. Marchetti, and G. Vailati, 1980. Influence of water hardness on cadmium toxicity to *Salmo gairdneri* Rich. *Water Research, 14:*1421–1426.

Carter, E.M. and J.J. Hubert, 1984. A growth-curve model approach to multivariate quantal bioassay. *Biometrics, 40:*699–700.

Caux, P.Y. and D.R.J. Moore, 1997. A spreadsheet program for estimating low toxic effects. *Environ. Toxicol. Chem., 16:*802–806.

CCREM [Canadian Council of Resource and Environment Ministers], 1987. *Canadian water quality guidelines.* CCREM, Task Force on Water Quality Guidelines. Environment Canada, Ottawa, Ont.

CETIS, 2001. Comprehensive Environmental Toxicity Information System. Tidepool Scientific Software, McKinleyville, Calif. 95521 [Program on disk and printed *User's Guide*.]

Chapman, P.M., 1996. Alternatives to the NOEC based on regression analysis. Discussion paper, Annex 7, OECD Workshop on Statistical Analysis of Aquatic Ecotoxicity Data, Braunschweig, Germany, Oct. 15-17, 1996, 5 p. Organisation for Economic Cooperation and Development, Paris. [Appendix A in Moore, 1996.]

Chapman, P.F., M. Crane, J. Wiles, F. Noppert, and E. McIndoe, 1996a. Asking the right questions: ecotoxicology and statistics. SETAC-Europe, Brussels [Society of Environmental Toxicology and Chemistry]. Report of a workshop held at Royal Holloway Univ. of London, Egham, Surrey, U.K., 26-27 April, 1995.

Chapman, P.M., R.S. Caldwell, and P.F. Chapman, 1996b. A warning: NOECs are inappropriate for regulatory use. *Environ. Toxicol. Chem., 15:*77–79.

Christensen, E.R., 1984. Dose-response functions in aquatic toxicity testing and the Weibull model. *Wat. Res., 18:* 213–221.

Christensen, E.R. and N. Nyholm, 1984. Ecological assays with algae: Weibull dose-response curves. *Env. Sci. Technol., 19:*713–718

Cochran, W.G. G.M. Cox, 1957. *Experimental designs.* 2nd ed. Wiley, New York, N.Y. 611 p.

Cohen, J., 1964. Psychological time. *Scientific Amer., 211, No. 5:*117–118.

Collett, D., 1991. *Modelling binary data*. Chapman & Hall, London. 369 p.

Crane, M. and E. Godolphin, 2000. Statistical analysis of effluent bioassays. Environment Agency, Bristol, U.K. Research and Development Tech. Rept E19.

Crane, M. and M.C. Newman, 2000. What level of effect is a no observed effect? *Environ. Toxicol. Chem., 19:*516–519.

Crane, M., M.C. Newman, P.F. Chapman, and J. Fenlon, 2002. *Risk assessment with time to event models.* Lewis Publishers/CRC Press, Boca Raton, Fla., 302 p.

Critchlow, D.E. and M.A. Fligner, 1991. On distribution-free multiple comparisons in the one-way analysis of variance. *Comm. Stat. Theory Methods, 20:*127–139.

D'Agostino, R.B., 1986. Tests for the normal distribution. p. 367–420, In: Goodness-of-fit techniques. R.B. D'Agostino and M.A. Stephens (eds.), Marcel Dekker Inc., New York, N.Y.

Damico, J.A. and D.A. Wolfe, 1987. Extended tables of the exact distribution of a rank statistic for treatment versus control multiple comparisons in one-way layout designs. *Comm. Stat. Theory Methods, 18:*3327–3353.

Davis, J.M. and D.J. Svendsgaard, 1990. U-shaped does-response curves: their occurrence and implications for risk assessment. *J. Toxicol. & Environ. Health, 30:*71–83.

Davis, R.B., A.J. Bailer, and J.T. Oris, 1998. Effects of organism allocation on toxicity test results. *Environ. Toxicol. Chem., 17:*928–931.

deBruijn, H.H.M. and M. Hof, 1997. How to measure no effect. Part IV: How acceptable is the ECx from an environmental policy point of view? *Environmetrics, 8:* 263–267.

Dixon, W.J. and F.J. Massey Jr., 1983. *Introduction to statistical analysis.* 4th ed. McGraw-Hill, New York, N.Y.

Dixon, P.M. and M.C. Newman, 1991. Analyzing toxicity data using statistical models for time-to-death: an introduction. p. 207–242, In: *Metal ecotoxicology, concepts and applications*. M.C. Newman and A.W. McIntosh (eds). Lewis Publishers, Inc., Chelsea, Mich., 399 p.

Dobson, A.J., 2002. *An introduction to generalized linear models.* 2nd ed. Chapman & Hall/CRC, Boca Raton, Fla. and London, U.K. 240 p.

Doe, K.G., 1994. Comments on the minutes of the Toxicological Statistics Advisory Group Meeting in Quebec City. Memorandum to J.A. Miller, Technology Development Branch, July 28, 1994. [K.G. Doe, Head, Toxicology Section, Environment Canada, Dartmouth, N.S.]

Douglas, M.T., D.O. Chanter, I.B. Pell, and G.M. Burney, 1986. A proposal for the reduction of animal numbers required for the acute toxicity to fish test (LC_{50} determination). *Aquat. Toxicol., 8:*243–249.

Draper, N.R. and J.A. John, 1981. Influential observations and outliers in regression. *Technometrics, 23:*21–26.

Draper, N.R. and H. Smith, 1981. *Applied regression analysis.* 2nd ed. Wiley, New York, N.Y. 709 p.

Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc., 50:*1096–1121.

———. 1964. New tables for multiple comparisons with a control. *Biometrics, 20:*482–491.

Dunnett, C.W. and A.C. Tamhane, 1998. New multiple test procedures for dose finding. *J. Biopharmaceut. Stat., 8:* 353–366.

Du Nouy, L., 1936. *Biological time.* Methuen, London. 180 p.

EC [Environment Canada], 1990a. Biological test method: Acute lethality test using rainbow trout. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/9. (with 1996 amendments).

———. 1990b. Biological test method: Acute lethality test using threespine stickleback *Gasterosteus aculeatus*. Environmental Protection Series, Ottawa, Ont., Rept. EPS 1/RM/10 (with 2002 amendments).

———. 1990c. Biological test method: Acute lethality test using *Daphnia* sp. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/11 (with 1996 amendments).

———. 1990d. Guidance document on control of toxicity test precision using reference toxicants. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/12.

———. 1992a. Biological test method: Test of reproduction and survival using the cladoceran *Ceriodaphnia dubia*. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/21 (with 1997 amendments).

———. 1992b. Biological test method: Test of larval growth and survival using fathead minnows. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/22. (with 1997 amendments).

———. 1992c. Biological test method: toxicity test using luminescent bacteria *(Photobacterium phosphoreum)*. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/24.

———. 1992d. Biological test method: growth inhibition test using the freshwater alga *Selenastrum capricornutum*. Environmental Protection Series, Ottawa, Ont., Rept. EPS 1/RM/25 (with 1997 amendments).

———. 1992e. Biological test method: acute test for sediment toxicity using marine or estuarine amphipods. Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/26 (with 1998 amendments).

———. 1992f. Biological test method: fertilization assay using echinoids (sea urchins and sand dollars). Environmental Protection Series, Ottawa, Ont., Rept EPS 1/RM/27 (with 1997 amendments).

———. 1992g. Fertilization assay with echinoids: interlaboratory evaluation of test options. EC, Conservation and Protection, Technol. Development Branch, Ottawa, Ont. Unpub. Rept. 45 p. + app.

———. 1994. Guidance document for the collection and preparation of sediments for physicochemical characterization and biological testing, Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/29, 144 p.

———. 1997a. Biological test method: Test for growth and survival in sediment using larvae of freshwater midges *(Chironomus tentans* or Chironomus riparius), Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/32.

———. 1997b. Biological test method: Test for survival and growth in sediment using the freshwater amphipod *Hyalella azteca,* Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/33.

———. 1998a. Biological test method: Toxicity tests using early life stages of rainbow trout. 2nd ed. Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/28.

———. 1998b. Biological test method: Reference method for determining acute lethality of sediment to marine or estuarine amphipods, Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/35.

———. 1999a. Guidance document on application and interpretation of single-species tests in environmental toxicology, Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/34.

———. 1999b. Biological test method: Test for measuring the inhibition of growth using the freshwater macrophyte *Lemna minor,* Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/37.

———. 2000a. Biological test method: Reference method for determining acute lethality of effluents to rainbow trout. 2nd ed. Method Development and Applications Centre, Ottawa, Ont., Rept EPS 1/RM/13.

———. 2000b. Biological test method: Reference method for determining acute lethality of effluents to *Daphnia magna*. 2nd ed. Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/14.

———. 2001a. Biological test method: test for survival and growth in sediment using spionid polychaete worms *(Polydora cornuta)*, Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/41.

———. 2001b. Revised procedures for adjusting salinity of effluent samples for marine sublethal toxicity testing conducted under Environmental Effects Monitoring (EEM) programs, Method Development and Applications Section, Ottawa, Ont. Unnumbered Rept, October 2001, 9 p.

———. 2002a. Biological test method: Reference method for determining the toxicity of sediment using luminescent bacteria (Vibrio fischeri), Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/42.

———. 2002b. Metal mining guidance document for aquatic environmental effects monitoring. Environmental Conservation Service, Ottawa, Ont.

———. 2004a. Biological test method: Tests for toxicity of contaminated soil to earthworms (*Eisenia andrei, Eisenia fetida,* or *Lumbricus terrestris,* Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/43.

———. 2004b. Biological test method: Test for measuring emergence and growth of terrestrial plants exposed to contaminants in soil, Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/45.

———. 2005. Biological test method: Test for measuring survival and reproduction effects in springtails (*Onychiurus folsomi* and *Folsomia candida*). Method Development and Applications Section, Ottawa, Ont., Rept EPS 1/RM/47.

Edwards, D. and J.J. Berry, 1987. The efficiency of simulation-based multiple comparisons. *Biometrics, 43:* 913–926.

Efron, B., 1982. The jackknife, the bootstrap and other resampling plans. Soc. Indust. Appld. Math, Philadelphia, Pa., CMMS 38.

Finney, D.J., 1971. *Probit analysis.* 3rd ed. Cambridge University Press, London. 333 p.

———. 1978. *Statistical method in biological assay*. 3rd ed. Charles Griffin & Co. Ltd, London. 508 p.

Finney, D.J., R. Latscha, B.M. Bennett, and P. Hsu, 1963. *Tables for testing significance in a 2 × 2 contingency table.* Published for the Biometrika Trustees, at the University Press, Cambridge, 102 p.

Fleiss, J.L., 1981. *Statistical methods for rates and proportions.* 2nd John Wiley & Sons, Toronto. 321 p.

Fligner, M.A. and D.A. Wolfe, 1982. Distribution-free tests for comparing several treatments with a control. *Stat. Neer., 36:*119–127.

Fry, F.E.J. 1947. Effects of the environment on animal activity. Univ. Toronto Studies, Biol. Series no. 55, *Publ. Ont. Fish. Res. Lab., 68:*1–62

Gad, S.C., 1999. *Statistics and experimental design for toxicologists*. CRC Press, Boca Raton, Fla. 437 p.

Gaddum, J.H., 1953. Bioassays and mathematics. *Pharmacol. Rev., 5:*87–134.

Gelber, R.D., P.T. Lavin, C.R. Mehta, and D.A. Schoenfeld, 1985. Statistical analysis. p. 110–123, In: *Fundamentals of aquatic toxicology. Methods and applications.*, G.M. Rand and S.R. Petrocelli (eds.), Hemisphere Publishing Corporation, Washington, D.C.

Grothe, D.R., K.L. Dickson, and D.K. Reed-Judkins, 1996. Whole effluent toxicity testing: An evaluation of methods and prediction of receiving system impacts. SETAC Press (Soc. Environmental Toxicol. and Chemistry] Pensacola, Fla. 346 p.

Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples. *Technometrics, 11:*1–21.

Hamilton, M.A., 1979. Robust estimates of the ED50. *J. Amer. Stat. Assoc., 74:*344–354.

———. 1980. Inference about the ED50 using the trimmed Spearman-Kärber procedure -- a Monte Carlo investigation. *Commun. Statist. Simula. Computa. B. 9(3):*235–254.

———. 1986. Statistical analysis of the cladoceran reproductivity test. *Environ. Toxicol. Chem., 5:*205–212.

Hamilton, M.A., R.C. Russo, and R.V. Thurston, 1977. Trimmed Spearman-Kärber method for estimating median lethal concentrations in toxicity bioassays. *Environ. Sci. Technol.11:*714–719. Correction. Same journal *12:*417.

Härdle, W., 1991. *Smoothing techniques with implementation in S.* Springer-Verlag, New York.

Hastie, T. and R. Tibshirani, 1990. *Generalized additive models*. Chapman and Hall, London.

Hayter, A.J. and G. Stone, 1991. Distribution-free multiple comparisons for monotonically ordered treatment effects. *Austral. J. Stat., 33:*335–346.

Heming, T.A., S. Arvind, and K. Yogesh, 1989. Time-toxicity relationships in fish exposed to the organochlorine pesticide methoxychlor. *Environ. Toxicol. Chem., 8:* 923–932.

Hewlett, P.S. and R.L. Plackett, 1979. *The interpretation of quantal responses in biology*. University Park Press, Baltimore, Maryland. 82 p.

Hochberg, Y. and A.C. Tamhane, 1987. *Multiple comparison procedures.* J. Wiley and Sons, New York, N.Y.

Hodson, P.V., C.W. Ross, A.J. Niimi, and D.J. Spry, 1977. Statistical considerations in planning aquatic bioassays. p. 15–31, In: *Proc. 3rd Aquatic Toxicity Workshop*, Halifax, N.S., Nov. 2–3, 1976. Environment Canada, Environmental Protection Service, Tech. Rpt No. EPS-5-AR-77-1, Halifax, Nova Scotia.

Hoekstra, J.A., 1989. Estimation of the ED50; letter to the editor. *Biometrics, 45:*337–338.

Hoekstra, J.A. and P.H. Van Ewijk, 1993. Alternatives for the no-observed-effect level. *Environ. Toxicol. Chem., 12:*187–194.

Hollander, M. and D.A. Wolfe, 1999. *Nonparametric statistical methods.* J. Wiley and Sons, New York, N.Y. 787 p.

Hong, W-H., P.G. Meier, and R.A Deininger, 1988. Determination of dose-time response relationships from long-term acute toxicity test data. *Environ. Toxicol. Chem., 7:*221–226.

Horness, B.H., D.P. Lomax, L.L. Johnson, M.S. Myers, S.M. Pierce, and T.K. Collier, 1998. Sediment quality thresholds: estimates from hockey stick regression of liver lesion prevalence in English sole *(Pleuronectes vetulus)*. *Environ. Toxicol. Chem., 17:*872–882.

Hosmer, D.W. and S. Lemeshow, 2000. *Applied logistic regression.* 2nd ed. Wiley-Interscience, New York, N.Y.

Hubert, J.J., 1984. *Bioassay*. 2nd ed. Kendall/Hunt Pub. Co., Dubuque, Iowa, 180 p.

———. 1992. *Bioassay*. 3rd ed. Kendall/Hunt Pub. Co., Dubuque, Iowa, 198 p.

———. 1987. PROBIT2: A microcomputer program for probit analysis. Dept Mathematics and Statistics, Univ. of Guelph, Guelph, Ont. N1G 2W1.

Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monog., 54:*187–2112.

ISO [International Organization for Standardization], 1999. Water quality -- Guidlines for alagal growth inhibition tests with poorly soluble materials, volatile compounds, metals and waste water. ISO, Geneva. ISO 14442, 14 p.

Jackman, P. and K. Doe, 2003. Evaluation of CETIS statistical software. Unnumbered report, Environment

Canada, Environmental Science Centre, Moncton, N.B. 46 p.

Jennrich, R.I. and R.H. Moore, 1975. Maximum likelihood estimation by means of nonlinear least squares. Proc. *Statistical Computing Section, Amer. Statistical Assoc., 70:*57–65.

Jensen, A.L., 1972. Standard error of LC_{50} and sample size in fish bioassays. *Water Res., 6:*85–89.

Jonckheere, A.R., 1954. A distribution free k-sample test against ordered alternatives. *Biometrika, 41:*133–145.

Kappenman, R.F., 1987. Nonparametric estimation of dose-response curves with application to ED50 estimation. *J. Stat. Com. Sim., 28:*1–13.

Kerr, D.R. and J.P. Meador, 1996. Modeling dose response using generalized linear models. *Environ. Toxicol. Chem., 15:*395–401.

Kooijman, S.A.L.M., 1996. An alternative for NOEC exists, but the standard model has to be abandoned first. *Oikos 75:* 310–316.

Kooijman, S.A.L.M. and J.J.M Bedaux, 1996. The analysis of aquatic toxicity data. VU Univ. Press, Vrije Universiteit, Amsterdam. 149 p. [Includes computer software disk DEBtox]

Kooijman, S.A.L.M., A.O. Hanstveit, and N. Nyholm, 1996. No-effect concentration in algal growth inhibition tests. *Water Res., 30:*1625–1632.

Koper, N., 1999. Nonlinear regression lecture for Vancouver workshop. Attachment Ra, 12 p., In: *Minutes/Proceedings of the Statistics Workshop for Toxicological Testing,* Pacific Environmental Science Centre (PESC), North Vancouver B.C., September 15–17th, 1999. Environment Canada, Pacific Environmental Science Centre, North Vancouver, B.C.

Kruskal, W.H. and W.A. Wallis, 1952. Use of ranks in one-criterion analysis of variance. *J. Amer. Statist. Assoc., 47:*583–621.

Lanno, R.P., G.L. Stephenson, and C.D. Wren, 1997. Applications of toxicity curves in assessing the toxicity of diazinon and pentachlorophenol to *Lumbricus terrestris* in natural soils. Soil Biology and Biochemistry 29: 689-692.

Lee, G., M.R. Ellersieck, F.L. Mayer,and G.F. Krause, 1995. Predicting chronic lethality of chemicals to fishes from acute toxicity test data: multifactor probit analysis. *Environ. Toxicol. Chem., 14:*345–349.

Levene, H., 1960. Robust tests for the equallity of variances. p. 278–292, In: *Contributions to probability and statistics.* I. Olkin (ed.), Stanford Univ. Press, Palo Alto, Calif.

Litchfield, J.T., 1949. A method for rapid graphic solution of time-percent effect curves. *Pharmacol. Exp. Ther., 97:*399–408.

Litchfield, J.T. and F. Wilcoxon, 1949. A simplified method of evaluating dose-effect experiments. *J. Pharmacol. Experimental Therapeutics, 96:*99–113.

Lloyd, Richard, 1992. *Pollution and freshwater fish. Fishing News Books* (Blackwell Scientific Publications Ltd), Oxford. 176 p.

Mallows, C.L., 1973. Some comments on C_p. *Technometrics, 12:*621–625.

Manly, B.F.J., 2000. *Statistics for environmental science and management.* CRC Press, Boca Raton, Fla. 336 p.

Marcus, A.H. and A.P. Holtzman, 1988. A robust statistical method for estimating effects concentrations in short-term fathead minnow toxicity tests. Battelle Washington Environmental Program Office, Washington, D.C. Report for USEPA Office of Water, Contract no. 69-03-3534, 39 p.

McCullagh, P. and J.A. Nelder, 1989. *Generalized linear models.* Chapman & Hall/CRC, Boca Raton, Fla. 532 p.

———. 1994. *Generalized linear models.* 2nd Chapman & Hall/CRC, Boca Raton, Fla., & London. 511 p.

McLeese, D.W., 1956. Effects of temperature, salinity and oxygen on the survival of the American lobster. *J. Fish. Res. Bd Canada, 13:*494–502.

Millard, S.P. and N.K. Neerchal, 2000. *Environmental statistics with S-Plus.* CRC Press, Boca Raton, Fla. 848 p.

Miller, R.G., 1981. *Simultaneous statistical inference.* Springer-Verlag, New York. 299 p.

———. 1986. *Beyond ANOVA, basics of applied statistics.* John Wiley & Sons, New York. [Cited through Newman, 1995.]

Miller, R.G. and J.W. Halpern, 1980. Robust estimators for quantal bioassay. *Biometrika, 67:*103–110.

Miller, J., R.P. Scroggins, and G.F. Atkinson, 1993. Toxicity endpoint determination statistics and computer

programs. Minutes of meeting of Statistical Advisory Group, Quebec City, October 20, 1993. Environment Canada, Technology Development Branch, Ottawa, Ont. 12 p. + attachments.

Moody, M. 2003. Research to assess potential improvements to Environment Canada's *Lemna minor* test method. Saskatchewan Research Council, Saskatoon, Sask., Pub. No. 11545-1C03. 69 p.

Moore D.R.J., 1996. OECD workshop on statistical analysis of aquatic ecotoxicity data. Summary report for Environment Canada. Unnumbered Report, Oct. 31, 1996, The Cadmus Group, Ottawa, Ont. 10 p. + appendices.

Moore, D.R.J. and P.-Y. Caux, 1997. Estimating low toxic effects. *Environ. Toxicol. Chem., 16:*794–801.

Moore, T.F., S.P. Canton, and M. Grimes, 2000. Investigating the incidence of type 1 errors for chronic whole effluent toxicity testing using *Ceriodaphnia dubia*. *Environ. Toxicol. Chem,. 19:*118–122.

Morissette, S., 2002. Le coût de l'incertitude en échantillonnage environnemental. Annexe C, In: *Environnement Canada. Guide d'échantillonnage des sédiments du Saint-Laurent pour les projets de dragage et de génie maritime. Vol. 1: Directives de planification.* Environnement Canada, Direction de la Protection de l'environnement, Région du Québec, Section innovation technologique et secteurs industriels. Rapport 106 p. [Available at http://www.slv2000.qc.ca.]

Müller, H.-G. and T. Schmitt, 1988. Kernel and probit estimates in quantal bioassay. *J. Amer. Stat. Assoc., 83:* 750–758.

Newman, M.C., 1995. *Quantitative methods in aquatic ecotoxicology.* Lewis Pub., Boca Raton, Fla. 426 p.

Newman, M.C. and M.S. Aplin, 1992. Enhancing toxicity data interpretation and prediction of ecological risk with survival time modeling: an illustration using sodium chloride toxicity to mosquitofish *Gambusia holbrooki*. *Aquatic Toxicol., 23:*85–96.

Noppert, F., N. van der Hoeven, and A. Leopold (eds), 1994. How to measure no effect. Towards a new measure of chronic toxicity in ecotoxicology. Workshop Rept, The Hague, The Netherlands, Sept. 9, 1994. The Netherlands Working Group on Statistics and Ecotoxicology. [Copies: BKH Consulting Engineers, P.O. box 5094, 2600 GB, Delft, The Netherlands, att. F. Noppert.]

Norberg-King, T.J., 1993. A linear interpolation method for sublethal toxicity: the Inhibition Concentration (ICp) approach (Version 2.0). USEPA, Duluth, Minn., Tech. Rept 03-93, National Effluent Toxicity Assessment Center, 25 p.

Nyholm, Niels, 2001. Personal communication. Comments on earlier draft of this document. Laboratory of Environ. Sciences and Ecology, Technical Univ. of Denmark, Lyngby, Denmark.

Nyholm, N., P.S. Sørensen, K.O. Kusk, and E.R. Christensen, 1992. Statistical treatment of data from microbial toxicity tests. *Environ. Toxicol. Chem., 11:* 157–167.

O'Brien, R.G., 1979. A general ANOVA method for robust tests of additive models for variances. *J. Amer. Stat. Assoc., 74:* 877–880.

OECD [Organisation for Economic Co-operation and Development], 1995. Guidance document for aquatic effects assessment. OECD, Paris. OECD Environment Monographs No. 92, 116 p.

———. 1997. Report of the final ring-test of the *Daphnia magna* reproduction test. OECD, Paris. OECD Environmental Health and Safety Publications, Series on Testing and Assessment No. 6.

———. 1998. Report of the OECD workshop on statistical analysis of aquatic toxicity data. OECD, Paris. OECD Environmental Health and Safety Publications, Series on Testing and Assessment No. 10, 133 p.

———. 2004. Draft guidance document on the statistical analysis of ecotoxicity data. OECD, Paris, Environmental Health and Safety Pub., Series on Testing and Assessment. 214 p. [available at www.oecd.org]

OMEE [Ontario Ministry of Environment and Energy], 1995. TOXSTATS. OMEE, Etobicoke, Ont. [Programs for estimating EC50 in a Windows format.]

Pack, S., 1993. A review of statistical data analysis and experimental design in OECD aquatic toxicology test guidelines. Shell Research Ltd., Sittingbourne Research Centre, Sittingbourne, Kent, U.K. 42. p.

———. 1998. A discussion of the NOEC/ANOVA approach to data analysis. Discussion paper, 9 p., In: *OECD, 1998. Report of the OECD workshop on statistical analysis of aquatic toxicity data. OECD, Paris.* OECD Environmental Health and Safety Publications, Series on Testing and Assessment No. 10, 133 p.

Paine, M.D., 1996. Letter to the Editor. Repeated measures designs. *Environ. Toxicol. Chem,. 13:*1439–1441.

effects monitoring (EEM) programs. *SETAC Globe, 3 (1):* 23–24. [Society of Environmental Toxicology and Chemistry, Pennsacola, Fla.]

Parmar, M.K.B. and D. Machin, 1995. *Survival analysis:Aa practical approach.* Wiley and Sons, New York.

Pickering, W., J. Lazorchak, and K. Winks, 1996. Subchronic sensitivity of one-, four-, and seven-day old fathead minnow *(Pimephales promelas)* larvae to five toxicants. *Environ. Toxicol. Chem., 15:*353–359.

Porebski, L.M. and J.M.Osborne, 1998. The application of a tiered testing approach to the management of dredged sediments for disposal at sea in Canada. *Chemistry and Ecology, 14:*197–214.

Rand, G.M. (ed.), 1995. *Fundamentals of aquatic toxicology: effects, environmental fate, and risk assessment.* 2nd ed. Taylor & Francis, Washington, D.C., 1125 p.

Rand, G.M. and S.R. Petrocelli (eds), 1985. *Fundamentals of aquatic toxicology.* Hemisphere Pub., Washington, D.C.

Ricker, W.E., 1958. Handbook of computations for biological statistics of fish populations. *Bull. Fish. Res. Bd Canada,* No. 119, 300 p.

Robertson, J.L., K.C. Smith, N.E. Savin, and J.L. Lavigne, 1984. Effects of dose selection and sample size on the precision of lethal dose estimates in dose-mortality regression. *J. Econ. Entomol., 77:*883–837.

Rosner, B., 1983. Percentage points for a generalized ESD many-outlier procedure. *Technometrics, 25:*165–172.

Rowe, D.W., J.B. Sprague, T.A. Heming, and I.T. Brown, 1983. Sublethal effects of treated liquid effluent from a petroleum refinery. II. Growth of rainbow trout. *Aquat. Toxicology, 3:*161–169.

Salsburg, D., 2001. *The lady tasting tea. How statistics revolutionized science in the twentieth century.* Henry Holt & Co., New York, N.Y. 340 p.

SAS [SAS Institute Inc.], 1988. SAS procedures guide, Release 6.03, *and* Additional SAS/STAT procedures, Release 6.03 (SAS Technical Report P-179). SAS Institute Inc., Cary, N.C. [http://www.sas.com]

———. 2000. SAS/STAT users guide, Version 9, SAS Institute Inc., Cary, N.C.

Scholze, M., W. Boedeker, M. Faust, T. Backhaus, R. Altenburger, and L.H. Grimme, 2001. A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. *Environ. Toxicol. Chem., 20:*448–457.

Schroeder, J. and R.P. Scroggins, 2001. Meeting notes. Discussion of comments on the fourth draft version of guidance document on statistical mehods to determine endpoints of toxicity tests. Sept. 27 and 28, 2001, Pacific Environmental Sciences Centre, North Vancouver, B.C.

Scott, D.W., 1992. *Multivariate density estimation. Theory, practice and visualization.* Wiley and Sons, New York, N.Y.

Scroggins, R.P., J.A. Miller, A.I. Borgmann, and J.B. Sprague, 2002. Sublethal toxicity findings by the pulp and paper industry for Cycles 1 and 2 of the environmental effects monitoring program. *Water Qual. Res. J. Canada, 37:(1):*21–48.

Searle, S.R., 1971. *Linear models.* John Wiley & Sons, New York, N.Y.

Sebaugh, J.L., 1998. Comparison of LC50 results from commonly used computer programs. p. 383-397, In: *Environmental toxicology and risk assessment: seventh volume.* E.E. Little, A.J. DeLonay, and B.M. Greenberg (eds), ASTM STP 1333, Amer. Soc. Testing and Materials, Philadelphia, Pa., 416 p.

Shapiro, S.S. and M.B. Wilk, 1965. An analysis of variance test for normality (complete samples). *Biometrika, 52:*591–611.

Shepard, M.P., 1955. Resistance and tolerance of young speckled trout *(Salvelinus fontinalis)* to oxygen lack, with special reference to low oxygen acclimation. *J. Fish. Res. Board Can., 12:*387–446.

Shirley, E., 1977. A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. *Biometrics, 33:*386–389.

Shukla, R., W. Wang, F. Fulk, C. Deng, and D. Denton, 2000. Bioequivalence approach for whole effluent toxicity testing. *Environ. Toxicol. Chem., 19:*169–174.

Slob, W., 2002. Dose-response modelling of continuous endpoints. *Toxicol. Sc., 66:*298–312.

Snedecor, G.W. and W.G. Cochran, 1980. *Statistical methods.* 7th ed. Iowa State Univ. Press, Ames, Iowa.

Sokal, R.R. and F.J. Rohlf, 1981. *Biometry.* W.H. Freeman and Co., San Francisco, Calif.

Sprague, J.B., 1964. Lethal concentrations of copper and zinc for young Atlantic salmon. *J. Fish. Res. Board Can. 21:*17–26.

————. 1969. Measurement of pollutant toxicity to fish -- I. Bioassay methods for acute toxicity. *Water Res., 3:* 793–821.

————. 1995. Factors that modify toxicity. p. 1012–1051, In: *Fundamentals of aquatic toxicology.* G.M. Rand (ed.). Taylor and Francis, Washington, D.C.

————. 1997. Review of methods for sublethal aquatic toxicity tests relevant to the Canadian metal-mining industry. Natural Resources Canada, Aquatic Effects Technol. Eval. Progr., Ottawa, Ont. AETE Project 1.2.1: 102 p.

Sprague, J.B. and A. Fogels, 1977. Watch the Y in bioassay. Proc. 3rd. Aquatic Toxicity Workshop, Halifax, N.S., Nov. 2–3, 1976. Environment Canada, Surveillance Rept. No. EPS-5-AR-77-1: 107-118.

SPSS, 1996. SPSS 6.1 for Windows. SPSS Inc., 233 South Wacker Drive, Chicago, Ill. 60606-5307, [Designed for Windows 3.1. Includes probit and logit regression.]

———. 2001. SPSS base 11.0 for Windows. SPSS Inc., 233 South Wacker Drive, Chicago, Ill. 60606-5307,[Probit and logit regression are in the "regression module" which can be added to the "base" statistical package.]

Steel, R.G.D., 1959. A multiple comparison rank sum test: treatments versus control. *Biometrics, 15:*560–572.

————. 1960. A rank-sum test for comparing all pairs of treatments. *Technometrics, 2:*197–611.

————. 1961. Some rank sum multiple comparison tests. *Biometrics, 17:*539–552.

Steel, R.G.D. and J.H. Torrie, 1980. *Principles and procedures of statistics.* 2nd ed. McGraw-Hill Book Co., New York.

Steel, R.G.D., J.H. Torrie, and D.A. Dickey, 1997. *Principles and procedures of statistics: a biometrical approach.* 3rd ed. McGraw-Hill Book Co., Boston. 666 p.

Stephan, C.E., 1977. Methods for calculating an LC_{50}. p. 65–84, In: *Aquatic toxicology and hazard evaluation.* F.L. Mayer and J.L. Hamelink (eds.), Amer. Soc. Testing and Materials, Philadelphia, Pa. ASTM STP No. 634.

Stephan, C.E., K.A. Busch, R. Smith, J. Burke, and R.W. Andrew, 1978. A computer program for calculating an LC50. [LC50.BAS] Provided courtesy of Dr. C.E. Stephan, U.S. Environmental Protection Agency, Duluth, Minn.

Stephan, C.E. and J.W. Rogers, 1985. Advantages of using regression analysis to calculate results of chronic toxicity endpoints. p. 328–338, In: *Aquatic toxicology and hazard assessment: Eighth symposium.* R.C. Bahner and D.J. Hansen (eds). Amer. Soc. Testing and Materials, Philadelphia, Pa. ASTM STP No. 891.

Stephan, C.E., D.I. Mount, D.J. Hansen, J.H. Gentile, G.A. Chapman, and W.A. Brungs, 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. USEPA, Office of Research and Development, Environmental Research Laboratories, Washington, D.C. [Available as NTIS PB 85-227049]

Stephenson, G.L., N. Koper, G.F. Atkinson, K.R. Solomon, and R.P. Scroggins, 2000. Use of nonlinear regression techniques for describing concentration-response relationships of plant species exposed to contaminated site soils. *Environ. Toxicol. Chem., 19:*2968–2981.

Suter, G.W. II, 1996. Abuse of hypothesis testing statistics in ecological risk assessment. *Human and Ecol. Risk Assess., 2:*331–347.

Suter, G.W. II, A.E. Rosen, E. Linder, and D.F. Parkhurst, 1987. Endpoints for responses of fish to chronic toxic exposures. *Environ. Toxicol. Chem., 6:*793–809.

Suter, G.W. II, B.W. Cornaby, C.T. Hadden, R.N. Hull, M. Stack, and F.A. Zafran, 1995. An approach for balancing health and ecological risks at hazardous waste sites. *Risk. Anal., 15:*221–231.

SYSTAT, 1990. SYSTAT: the system for statistics. SYSTAT Inc., Evanston, Ill. 677 p.

Thompson, W.R., 1947. Use of moving averages and interpolation to estimate median-effective dose. 1. Fundamental formulas, estimation of error, and relation to other methods. *Bact. Reviews, 11:*115–145.

Tattersfield, F. and H.M. Morris, 1924. An apparatus for testing the toxic values of contact insecticides under controlled conditions. *Bull. Entomological Res., 14:* 223–233.

TOXCALC. Version 5.0, 1994. Tidepool Scientific Software, McKinleyville, Calif. 95521. [Program on disk with printed user's manual.] [Superceded in 2001 by the package *CETIS, q.v.*] [http://members.aol.com/tidesoft/toxcalc]

TOXSTAT, 1996. Version 3.5. Lincoln Research Associates, Inc., P.O. Box 4276, Bisbee, Ariz., 85603, email danlra@msn.com. [Programs on disk with printed user's manual.]

Tukey, J.W., 1977. *Exploratory data analysis*. Addison-Wesley, Reading, Mass. 688 p.

USEPA [United States Environmental Protection Agency], 1991. Technical support document for water quality-based toxics control. USEPA, Office of Water, Washington, D.C., EPA/505/2-90-001.

———. 1994a. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms. 3rd ed. U.S. EPA, Environmental Monitoring Systems Laboratory, Cincinnati, Ohio, EPA 600/4-91-002.

———. 1994b. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to marine and estuarine organisms. 2nd ed. USEPA, Environmental Monitoring Systems Laboratory, Cincinnati, Ohio, EPA 600/4-91/003.

———. 1994c. Methods for measuring the toxicity and bioaccumulation of sediment-associated contaminants with freshwater invertebrates. USEPA, Duluth, Minn., EPA/600/R-94/024.

———. 1994d. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. USEPA, Office of Research and Development, Washington, D.C. EPA/600/R-94/025.

———. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. G.A. Chapman, D.L. Denton, and J.M. Lazorchak (eds.), USEPA, Office of Research and Development, Washington, D.C., EPA 600/R-95/136, 661 p.

———. 2000a. Method guidance and recommendations for whole efluent toxicity (WET) testing (40 CFR Part 136). USEPA, Office of Water, Washington, D.C., EPA 821-B-00-004. 60 p.

———. 2000b. Understanding and accounting for method variability in whole effluent toxicity applications under the national pollutant discharge elimination system program. USEPA, Office of Wastewater Management Washington, D.C., EPA/833/R-00/003.

USEPA and USACE [United States Environmental Protection Agency and United States Army Corps of Engineers], 1994. Evaluation of dredged material proposed for discharge in inland and near coastal waters. USEPA, Office of Science and Technology, Washington, D.C., EPA/000/0-93/000.

van der Hoeven, N., 1991. LC_{50} estimates and their confidence intervals derived for tests with only one concentration with partial effect. *Water Res., 25:*401–408

———. 1997. How to measure no effect. Part III: Statistical aspects of NOEC, ECx and NEC estimates. *Environmetrics, 8:*255–261.

van der Hoeven, N., F. Noppert, and A. Leopold, 1997. How to measure no effect. Part I: Towards a new measure of chronic toxicity in ecotoxicology. Introduction and workshop results. *Environmetrics, 8:*241–248.

Van Ewijk, P.H. and J.A. Hoekstra, 1993. Calculation of the EC50 and its confidence interval when subtoxic stimulus is present. *Ecotox. Env. Safety 25:*25–32.

Villeneuve, D.L., A.L. Blankenship, and J.P. Giesy, 2000. Derivation and application of relative potency estimates based on in vitro bioassay results. *Environ. Toxicol. Chem., 19:*2835–2843.

Wang, Q., D.L. Denton, and R. Shukla, 2000. Applications and statistical properties of minimum significant difference-based criterion testing in a toxicity testing program. *Environ. Toxicol. Chem., 19:*113–117.

Wang, S.C.D. and E.P. Smith, 2000. Adjusting for mortallity effects in chronic toxicity testing: mixture model approach. *Environ. Toxicol. Chem., 19:*204–209.

Wardlaw, A.C., 1985. *Practical statistics for experimental biologists.* John Wiley & Sons, Toronto, Ont..

Warren, C.E. 1971. *Biology and control of water pollution.* Saunders, Toronto, Ont. 434 p.

Wellek, S., 2002. *Testing statistical hypotheses of equivalence.* Chapman & Hall/CRC, Boca Raton, Fla. 290 p.

WEST, Inc. and D.D. Gulley, 1996. Toxstat® 3.5. Western EcoSystems Technology, Inc., Cheyenne, Wyo., U.S.A. [Computer software and instruction manual.]

Wilber, C.G., 1962. The biology of water toxicants in sublethal concentrations. p. 326–331, In: *Biological problems in water pollution. Third seminar.* C.M. Tarzwell (ed.), U.S. Public Health Service, Dept. Health, Education, and Welfare, R.A. Taft Sanitary Engnrng Center, Cincinnati, Ohio, P.H.S. pub. no. 999-WP-25.

Williams, D.A., 1971, A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics, 27:*103–117.

———. 1972. The comparison of several dose levels with a zero dose control. *Biometrics, 28:*519–531.

WSDOE [Washington State Dept of Ecology], 1998. Laboratory guidance and whole effluent toxicity test review criteria. WSDOE, Water Quality Program, Pub. no. WQ-R-95-80, 71 p. Olympia, Wash.

Zajdlik, B.A., 1996. An introduction to threshold modelling of non-quantal bioassay data. p. 89–96, In: *Proc. 22nd Ann. Aquat. Toxicity Workshop: October 2-4, 1995, St. Andrews, New Brunswick.* K. Haya and A.J. Niimi (eds.), Fisheries and Oceans, Can. Tech. Rept Fisheries and Aquatic Sc. No. 2093.

———. in preparation. Methods for statistically comparing EC50s and ICps. B. Zajdlik & Associates Inc., Rockwood, Ont.

Zajdlik, B.A., K.G. Doe, and L.M. Porebski, 2000. Report on biological toxicity tests using pollution gradient studies -- Sydney Harbour. Environment Canada, Environment Protection Service, Marine Environment Div., EPS 3/AT/2. 104 p.

Zajdlik, B.A., G. Gilron, P. Riebel, and G. Atkinson, 2001. The $500,000 fish. *SETAC Globe, 2 (1):* 28–30. [Society of Environmental Toxicology and Chemistry, Pennsacola, Fla.]

Zaleski, R.T., G.E. Bragin, M.J. Nicolich, W.R. Arnold, and A.L. Middleton, 1997. Comparison of growth endpoint estimation methods in EPA effluent short-term chronic testing guidelines. Poster PWA088, Soc. Environ. Toxicology and Chemistry, 18th Annual Meeting, San Francisco, Calif., 16-20 Nov. 1997.

Zar, J.H., 1974. *Biostatistical analysis.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

———. 1999. *Biostatistical analysis.* 4th ed. Prentice-Hall, Inc., Upper Saddle River, N.J.

Biological Test Methods and Supporting Guidance Documents Published by the Method Development and Applications Section, Environment Canada[1]

Title	Type of data	Date [2]	Amended
A. Generic (Universal) Biological Test Methods			
Acute lethality test using rainbow trout. [EPS 1/RM/9]	Quantal: acute mortality	July 1990 (1990a)	May 1996
Acute lethality test using threespine stickleback (*Gasterosteus aculeatus)*. [EPS 1/RM/10]	Quantal: acute mortality	July 1990 (1990b)	Mar. 2000
Acute lethality test using *Daphnia* spp. [EPS 1/RM/11]	Quantal: acute mortality	July 1990 (1990c)	May 1996
Reproduction and survival using the cladoceran *Ceriodaphnia dubia*. [EPS 1/RM/21]	Dual effect: mortality of adults and number of young	Feb. 1992 (1992a)	Nov. 1997
Larval growth and survival using fathead minnows. [EPS 1/RM/22]	Dual effect: mortality and weight of larvae	Feb. 1992 (1992b)	Nov. 1997
Luminescent bacteria (*Photobacterium phosphoreum*) [now *Vibrio fischeri*] [EPS 1/RM/24]	Quantitative: 50% inhibition of light production	Nov. 1992 (1992c)	—
Growth inhibition using the freshwater alga *Selenastrum capricornutum [*now *Pseudokirchneriella subcapitata]*. [EPS 1/RM/25]	Quantitative: specified % reduction in algal cells produced during 72 hours	Nov. 1992 (1992d)	Nov. 1997
Acute test for sediment toxicity using marine or estuarine amphipods. [EPS 1/RM/26]	Quantal: % survival, emergence from sediment, failed to rebury	Dec. 1992 (1992e)	Oct. 1998
Fertilization assay using echinoids (sea urchins and sand dollars). [EPS 1/RM/27]	Quantal: reduced success of fertilization	Dec. 1992 (1992f)	Nov. 1997
Early life stages of salmonid fish (rainbow trout). [EPS 1/RM/28, 2nd edition]	Quantal: nonviable embryos, alevins, or fry; fry mortality. Quantitative: fry weight. Describe delayed, abnormal development	July 1998 (1998a)	—
Survival and growth in sediment using the larvae of freshwater midges (*Chironomus tentans* or *Chironomus riparius*). [EPS 1/RM/32]	Dual effect: survival and weight of larvae	Dec. 1997 (1997a)	—
Survival and growth in sediment using the freshwater amphipod *Hyalella azteca*. [EPS 1/RM/33]	Dual effect: survival and weight	Dec. 1997 (1997b)	—

1 These documents may be purchased from Environmental Protection Publications, Environment Canada, Ottawa, Ont., K1A 0H3. For information or comments, contact the Chief, Biological Methods Division, Environmental Technology Centre, Environment Canada, Ottawa, Ont. K1A 0H3.

2 The Date column gives publication date and also coding to the present reference list (e.g., 1990a).

Title	Type of data	Date	Amended
Inhibition of growth using the freshwater macrophyte *Lemna minor*. [EPS 1/RM/37]	Dual effect: weight and decreased proliferation (number) of fronds	Mar. 1999 (1999b)	—
Survival and growth in sediment using spionid polychaete worms (*Polydora cornuta*). [EPS 1/RM/41]	Dual effect: survival and weight	Dec. 2001 (2001a)	—
Tests for toxicity of contaminated soil to earthworms (*Eisenia andrei, Eisenia fetida,* or *Lumbricus terrestris*). [EPS 1/RM/43]	Quantal: acute mortality; per cent avoidance. Dual effect: adult mortality, number and weight of young.	June 2004 (2004a)	—
Test for measuring emergence and growth of terrestrial plants exposed to contaminants in soil. [EPS 1/RM/45]	Dual effect: number of seedlings emerged, length and weight of shoots and roots	June 2004 (2004b)	—
Survival and reproduction effects in springtails *(Onychiurus folsomi* and *Folsomia candida)*. [EPS 1/RM/47]	Dual effect: survival of adults and number of young	Dec. 2005 (2005)	—
B. Reference Test Methods [3]			
Acute lethality of effluents to rainbow trout. [EPS 1/RM/13, 2nd edition]	Quantal: acute mortality	Dec. 2000 (2000a)	—
Acute lethality of effluents to *Daphnia magna*. [EPS 1/RM/14, 2nd edition]	Quantal: acute mortality	Dec. 2000 (2000b)	—
Acute lethality of sediment to marine or estuarine amphipods. [EPS 1/RM/35]	Quantal: acute survival	Dec. 1998 (1998b)	—
Toxicity of sediment using luminescent bacteria in a solid-phase test. [EPS 1/RM/42]	Quantitative: inhibition of light production	Apr. 2002 (2002a)	—
C. Supporting Guidance Documents			
Control of toxicity test precision using reference toxicants. [EPS 1/RM/12]		Aug. 1990 (1990d)	—
Collection and preparation of sediment for physicochemical characterization and biological testing. [EPS 1/RM/29]		Dec. 1994 (1994)	—
Measurement of toxicity test precision using control sediments spiked with a reference toxicant. [EPS 1/RM/30]		Sept. 1995	—
Application and interpretation of single-species tests in environmental toxicology. [EPS 1/RM/34]		Dec. 1999 (1999a)	—
Guidance document for testing the pathogenicity and toxicity of new microbial substances to aquatic and terrestrial organisms. [EPS 1/RM/44]		Mar. 2004 (2004d)	—
Guidance document on statistical methods for environmental toxicity tests. [EPS 1/RM/46] [This document].		Mar. 2005	—

[3] A *reference method* is defined as a specific biological method for performing a toxicity test, having a set of instructions and conditions which are described precisely in a written document. Reference tests are usually associated with requirements of specific regulations, unlike the multi-purpose use of generic ("universal") biological test methods of Environment Canada.

Members of the Inter-Governmental Environmental Toxicity Group (as of January 2004)

Federal, Environment Canada

C. Blaise
Centre St. Laurent
Montreal, Quebec

M. Bombardier
Environmental Technology Centre
Ottawa, Ontario

U. Borgmann
National Water Research Institute
Burlington, Ontario

J. Bruno
Pacific Environmental Science Centre
North Vancouver, British Columbia

C. Buday
Pacific Environmental Science Centre
North Vancouver, British Columbia

K. Doe
Atlantic Environmental Science Centre
Moncton, New Brunswick

G. Elliott
Environmental Protection Service
Edmonton, Alberta

F. Gagné
Centre St. Laurent
Montreal, Quebec

M. Harwood
Environmental Protection Service
Montreal, Quebec

D. Hughes
Atlantic Environmental Science Centre
Moncton, New Brunswick

P. Jackman
Atlantic Environmental Science Centre
Moncton, New Brunswick

N. Kruper
Environmental Protection Service
Edmonton, Alberta

M. Linssen
Pacific Environmental Science Centre
North Vancouver, British Columbia

D. MacGregor
Environmental Technology Centre
Ottawa, Ontario

L. Porebski
Marine Environment Branch
Gatineau, Quebec

J. Princz
Environmental Technology Centre
Ottawa, Ontario

G. Schroeder
Pacific Environmental Science Centre
North Vancouver, British Columbia

R.P. Scroggins
Environmental Technology Centre
Ottawa, Ontario

T. Steeves
Atlantic Environmental Science Centre
Moncton, New Brunswick

D. Taillefer
Marine Environment Branch
Gatineau, Quebec

S. Trottier
Centre St. Laurent
Montreal, Quebec

G. van Aggelen (Chairperson)
Pacific Environmental Science Centre
North Vancouver, British Columbia

B. Walker
Centre St. Laurent
Montreal, Quebec

P.G. Wells
Environmental Conservation Service
Dartmouth, Nova Scotia

Federal, Fisheries & Oceans Canada

R. Roy
Institut Maurice Lamontagne
Mont-Joli, Quebec

Federal, Natural Resources Canada

J. McGeer
Mineral Sciences Laboratory, CANMET
Ottawa, Ontario

B. Vigneault
Mineral Sciences Laboratory, CANMET
Ottawa, Ontario

J. Beyak
Mineral Sciences Laboratory, CANMET
Ottawa, Ontario

Provincial

C. Bastien
Ministère de l'Environnement du Québec
Ste. Foy, Quebec

B. Bayer
Manitoba Environment
Winnipeg, Manitoba

M. Mueller
Ontario Ministry of Environment
Rexdale, Ontario

D. Poirier
Ontario Ministry of Environment
Rexdale, Ontario

J. Schroeder
Ontario Ministry of Environment
Rexdale, Ontario

T. Watson-Leung
Ontario Ministry of Environment
Rexdale, Ontario

Environment Canada Regional and Headquarters Offices

Headquarters
351 St. Joseph Boulevard
Place Vincent Massey,
Gatineau, Quebec
K1A 0H3

Ontario Region
4905 Dufferin St., 2nd Floor
Downsview, Ontario
M3H 5T4

Atlantic Region
15th Floor, Queen Square
45 Alderney Drive
Dartmouth, Nova Scotia
B2Y 2N6

Prairie and Northern Region
Room 210, Twin Atria No. 2
4999 -- 98th Avenue
Edmonton, Alberta
T6B 2X3

Québec Region
105 rue McGill
8ième étage
Montreal, Quebec
H2Y 2E7

Pacific and Yukon Region
401 Burrard Street
Vancouver, British Columbia
V6C 3S5

Calculations Using Arithmetic and Logarithmic Concentrations

D.1 Example: Comparing Means

The table shows the divergence between medians, arithmetic means, and geometric/logarithmic means for four hypothetical sets of numbers, and the columns might represent the EC50s for replicate tests. The first column represents "good" data, with results being fairly similar. The second column has a slightly divergent concentration at the high end of the set. The third column has a most unlikely high concentration. The fourth column has a wildly unlikely outlying concentration. It is assumed for the purpose of illustration, that there is no reason to reject any concentration. Any general principle which is derived from the extreme examples, would also apply to ordinary sets of data from toxicity laboratories.

	"Good" data	Divergent value	Unlikely value	Weird data
	10	10	10	10
	12	12	12	12
	14	14	14	14
	16	16	16	16
	18	18	18	18
	22	28	100	1000
Median	15	15	15	15
Arithmetic mean	15.3	16.3	28	178
Geometric mean	14.6	15.4	19	28

For the "good" set of concentrations in the first column, the three measures of central tendency are essentially the same, as would be expected for regular data. For the second, third, and fourth columns, the median remains the same, because it does not consider the numerical value of the highest item. The median could often be a good choice for expressing central tendency of a skewed distribution. Indeed, in estimating an EC50, the foundation of the endpoint is the quantal effect on the *median* tested organism. However, elsewhere in toxicology, the median has seldom been used for quantitative things like concentration, as investigators employ instead, an average that makes use of the numerical values. In these examples, the median fails to indicate that a high value is aberrant; even if both of the uppermost values in the set were aberrantly high, the median would show no change.

The arithmetic mean for the second column is about 6% higher than the geometric one, a difference that is noticeable but not of major importance.

For the third "unlikely" set of concentrations, the arithmetic mean is higher by a factor of almost 1.5, an appreciable difference. The geometric mean tends to minimize the effect of the outlier, and is more representative of the other five closely sequenced concentrations.

In the weird example, the arithmetic mean is 5.4 times higher than the geometric mean, and is not at all representative of most of the values in the series. The geometric mean is, at least, the same order of magnitude as the five similar concentrations.

Normally, the outliers in the two right-hand columns might be rejected by statistical testing, but that is not the point of this example. In the two extreme examples, the geometric mean clearly gives a more robust defence than does the arithmetic average, against the unusual high concentrations, and would seem to give a better representation of the probable average toxicity. The principle having been ascertained, it would also extend to "good" data-sets. The geometric mean should give a more dependable representation of mean values. Readers might wish to put together other examples.

D.2 Example: Probit Regressions

The following table gives calculations of EC50s by probit regression. The four examples are for the sets of data listed in Table 2 of Section 4.4. The estimates of EC50 using log concentrations are those obtained in most computer programs, which automatically use log concentration for calculation. The estimates of EC50 using arithmetic concentration were made using the program TOXSTAT 3.5, bypassing the use of logarithms. (It would not be difficult to make that mistake with the program, and never realize it, which is a good reason for checking estimates by hand plots.)

	EC50s (and confidence limits) for four example sets of data			
	A	B	C	D
With arithmetic concentration	6.3 (4.9–7.7)	20.6 (14.3–26.9)	15.6 (11.4–19.5)	32.5 (17.6–47.4)
With log concentration	5.6 (4.4–7.2)	16.8 (12.1–23.3)	12.8 (9.4–17.6)	26.5 (13.3–53.1)
Ratio, arithmetic/log	1.12	1.23	1.22	1.23

The EC50s calculated with arithmetic concentrations average 1.2 times the proper values. This is an appreciable error and should be avoided. Most of the confidence limits are also raised to generally higher ranges. Section 4.4 and Table 2 indicate that when proper logarithmic concentrations are used, the TOXSTAT endpoints are in substantial agreement with calculations by other programs.

Randomization

Randomization is part of assigning test organisms to containers and concentrations, and of assigning containers into an array.

E.1 Random Numbers for Allocating Organisms to Containers

Randomization of organisms into containers is not required in all methods published by Environment Canada. It was decided that in some tests, the procedures could lead to operator errors which would be more serious. Subsequent randomization of containers or concentrations was deemed sufficient to avoid bias in the test and its results.

However, if test organisms can be handled as discrete individuals (e.g., fish, as used here), and if they are to be counted into test containers, it is always advantageous to do so randomly. Any convenient method could be used, such as drawing slips of paper from a hat, the slips marked with concentrations. Most computers will generate random numbers. Another convenient way is offered by USEPA (1995), using a table of random numbers, and is repeated here (Table E.1).

First, a series of two-digit numbers are assigned to the various test concentrations, setting up in tabular form as shown immediately below. Several two-digit numbers should be used for each concentration, so that a later step "uses up" the numbers in a table of random numbers. The value 00 is not used, and in this particular tabulation, no number greater than 30 is used.

Assigned numbers					Test concentration
01	07	13	19	25	Control
02	08	14	20	26	0.5% effluent
03	09	15	21	27	1% effluent
04	10	16	22	28	2.5% effluent
05	11	17	23	29	5% effluent
06	12	18	24	30	10% effluent

Now, going to a table of random numbers such as Table E.1, any row and column may be picked to start (for example, row 3 of column 6, which has a value of 19). This is considered the first fish that happens to be caught from the stock tank, and from the tabulation above, it is assigned to the control.

Returning to Table E.1 the second fish is picked by moving horizontally to the right; the numbers 64, 50, and 93 are ignored because they are higher than those used in the tabulation. The second fish gets number 03, which assigns it to 1% effluent. The process continues across the row in Table E.1, then across the next row downwards, until the tanks are filled, say with 10 fish each. It is necessary to keep track of the assignments, so that each tank gets its full complement of fish, but no extras. If a number comes up that would overfill a container, the number is ignored. If one person is doing the exercise, it is easiest to do all the selection of numbers on paper, then catch and distribute fish.

Randomization must be done anew for each test or set of tests. It is not suitable to set up a pattern of randomization and use it over again.

Table E.1 Two-digit random numbers. From Dixon and Massey (1983; as used by USEPA, 1995).

10 09 73 25 33	76 52 01 35 86	34 67 35 43 76	80 95 90 91 17	39 29 27 49 45
37 54 20 48 05	64 89 47 42 96	24 80 52 40 37	20 63 61 04 02	00 82 29 16 65
08 42 26 89 53	19 64 50 93 03	23 20 90 25 60	15 95 33 47 64	35 08 03 36 06
99 01 90 25 29	09 37 67 07 15	38 31 13 11 65	88 67 67 43 97	04 43 62 76 59
12 80 79 99 70	80 15 73 61 47	64 03 23 66 53	98 95 11 68 77	12 27 17 68 33
66 06 57 47 17	34 07 27 68 50	36 69 73 61 70	65 81 33 98 85	11 19 92 91 70
31 06 01 08 05	45 57 18 24 06	35 30 34 26 14	86 79 90 74 39	23 40 30 97 32
85 26 97 76 02	02 05 16 56 92	68 66 57 48 18	73 05 38 52 47	18 62 38 85 79
63 57 33 21 35	05 32 54 70 48	90 55 35 75 48	28 46 82 87 09	83 49 12 56 24
73 79 64 57 53	03 52 96 47 78	35 80 83 42 82	60 93 52 03 44	35 27 38 84 35
98 52 01 77 67	14 90 56 86 07	22 10 94 05 58	60 97 09 34 33	50 50 07 39 98
11 80 50 54 31	39 80 82 77 32	50 72 56 82 48	29 40 52 42 01	52 77 56 78 51
83 45 29 96 34	06 28 89 80 83	13 74 67 00 78	18 47 54 06 10	68 71 17 78 17
88 68 54 02 00	86 50 75 84 01	36 76 66 79 51	90 36 47 64 93	29 60 91 10 62
99 59 46 73 48	87 51 76 49 69	91 82 60 89 28	93 78 56 13 68	23 47 83 41 13
65 48 11 76 74	17 46 85 09 50	58 04 77 69 74	73 03 95 71 86	40 21 81 65 44
80 12 43 56 35	17 72 70 80 15	45 31 82 23 74	21 11 57 82 53	14 38 55 37 63
74 35 09 98 17	77 40 27 72 14	43 23 60 02 10	45 52 16 42 37	96 28 60 26 55
69 91 62 68 03	66 25 22 91 48	36 93 68 72 03	76 62 11 39 90	94 40 05 64 18
09 89 32 05 05	14 22 56 85 14	46 42 75 67 88	96 29 77 88 22	54 38 21 45 98
91 49 91 45 23	68 47 92 76 86	46 16 28 35 54	94 75 08 99 23	37 08 92 00 48
80 33 69 45 98	26 94 03 68 58	70 29 73 41 35	53 14 03 33 40	42 05 08 23 41
44 10 48 19 49	85 15 74 79 54	32 97 92 65 75	57 60 04 08 81	22 22 20 64 13
12 55 07 37 42	11 10 00 20 40	12 86 07 46 97	96 64 48 94 39	28 70 72 58 15
63 60 64 93 29	16 50 53 44 84	40 21 95 25 63	43 65 17 70 82	07 20 73 17 90
61 19 69 04 46	26 45 74 77 74	51 92 43 37 29	65 39 45 95 93	42 58 26 05 27
15 47 44 52 66	95 27 07 99 53	59 36 78 38 48	82 39 61 01 18	33 21 15 94 66
94 55 72 85 73	67 89 75 43 87	54 62 24 44 31	91 19 04 25 92	92 92 74 59 73
42 48 11 62 13	97 34 40 87 21	16 86 84 87 67	03 07 11 20 59	25 70 14 66 70
23 52 37 83 17	73 20 88 98 37	68 93 59 14 16	26 25 22 96 63	05 52 28 25 62
04 49 35 24 94	75 24 63 38 24	45 86 25 10 25	61 96 27 93 35	65 33 71 24 72
00 54 99 76 54	64 05 18 81 59	96 11 96 38 96	54 69 28 23 91	23 28 72 95 29
35 96 31 53 07	26 89 80 93 45	33 35 13 54 62	77 97 45 00 24	90 10 33 93 33
59 80 80 83 91	45 42 72 68 42	83 60 94 97 00	13 02 12 48 92	78 56 52 01 06
46 05 88 52 36	01 39 09 22 86	77 28 14 40 77	93 91 08 36 47	70 61 74 29 41
32 17 90 05 97	87 37 92 52 41	05 56 70 70 07	86 74 31 71 57	85 39 41 18 38
69 23 46 14 06	20 11 74 52 04	15 95 66 00 00	18 74 39 24 23	97 11 89 63 38
19 56 54 14 30	01 75 87 53 79	40 41 92 15 85	66 67 43 68 06	84 96 28 52 07
45 15 51 49 38	19 47 60 72 46	43 66 79 45 43	59 04 79 00 33	20 82 66 95 41
94 86 43 19 94	36 16 81 08 51	34 88 88 15 53	01 54 03 54 56	05 01 45 11 76
98 08 62 48 26	45 24 02 84 04	44 99 90 88 96	39 09 47 34 07	35 44 13 18 80
33 18 51 62 32	41 94 15 09 49	89 43 54 85 81	88 69 54 19 94	37 54 87 30 43
80 95 10 04 06	96 38 27 07 74	20 15 12 33 87	25 01 62 52 98	94 62 46 11 71
79 75 24 91 40	71 96 12 82 96	69 86 10 25 91	74 85 22 05 39	00 38 75 95 79
18 63 33 25 37	98 14 50 65 71	31 01 02 46 74	05 45 56 14 27	77 93 89 19 36
74 02 94 39 02	77 55 73 22 70	97 79 01 71 19	52 52 75 80 21	80 81 45 17 48
54 17 84 56 11	80 99 33 71 43	05 33 51 29 69	56 12 71 92 55	36 04 09 03 24
11 66 44 98 83	52 07 98 48 27	59 38 17 15 39	09 97 33 34 40	88 46 12 33 56
48 32 47 79 28	31 24 96 47 10	02 29 53 68 70	32 30 75 75 46	15 02 00 99 94
69 07 49 41 38	87 63 79 19 76	35 58 40 44 01	10 51 82 16 15	01 84 87 69 38

E.2 Random Numbers for Allocating Positions of Chambers

Location of containers is important in a test. If the containers are in a line, it might be that one end was near a window, with strong direct lighting of some containers, resulting in stress, attempted avoidance reactions, or better growth, depending on the type of organism. One end of the series might be warmer than the other, a particular risk if the test was done in an incubator. One end of a series might be near a door and human traffic could startle the organisms. Such influences could affect results, and there could be other unexpected and unknown influences. Investigators should attempt to eliminate or minimize such influences, but there might be unrecognized factors. The way to eradicate any systematic uncontrolled variables is to randomize the positions of test chambers.

In the previous example, there were five concentrations and a control, and they might be placed in a row for convenience. Their positions could be randomized by picking numbers out of a hat or, as before, from Table E.1 or from a simpler table in a mathematical book.

If there were more containers to deal with, say five replicates of each concentration, the investigator might wish to have a 6 by 5 layout on the lab bench. (It could be any layout, the same procedure would apply.) The same randomizing process could be used as previously outlined.

- A tabulation is made and numbered to represent the 30 positions in the 5 by 6 configuration. Numbers from 01 to 30 are written into the table.

The tabulation of positions for test containers could appear like this one.

01	02	03	04	05	06
07	08	09	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30

- Enter Table E.1 at any point and read the number; it will represent the first replicate of the control. (e.g., the 11th column of Table E.1, 3rd row down. The number is 23, so the first control replicate goes to position 23 on the tabulation, near the lower right corner.)

- Move to the next number to the right in Table E.1, this number is for the second replicate of the control. (It is number 20, so this replicate will be positioned near the lower left corner of the tabulation.)

- Continue the process until all replicates of all concentrations have been assigned a position. If a number comes up a second time, ignore it. If a number over 30 comes up, ignore it. (The next number in our example would be 90, which does not apply to any position, so it is ignored, and the following number 25 becomes the position of the 3rd replicate of the control.)

In the tabulation shown above for numerals one to thirty, two or three numbers could have been listed for each position if desired, in order to better "use up" the numbers of Table E.1, and not draw so many numbers that must be ignored. Using only one number in each position simply means that the investigator will draw mostly "blanks" from Table E.1, i.e., most numbers taken from the table will not apply to any position and will be ignored.

Setting up the positions in this way would also be perfect preparation for a "blind" set of observations. The positions of the various replicates would be recorded, but unknown to the observer during the test. The observer

could not introduce bias from knowledge of the concentrations. After the end of the test, the observations would be assigned to their proper replicates and concentrations.

Possible exception. Testing volatile substances might be one of the few situations in which randomization of test containers in an open array would not be appropriate. The volatile toxicant might escape from containers with high concentrations and reach other containers. In particular, it might contaminate the control and cause anomalous control effects. Although this could certainly affect statistical analysis, the remedy lies in another field, that of proper laboratory facilities designed for testing volatile materials. Such a situation would require sealed containers, separate venting, or some such arrangement. Randomization of treatments would still be an objective.

Calculating the Mean and Limits for a Warning Chart

For convenience, Figure 2 is repeated as Figure F.1 in this appendix. Data from the figure will be used to show the steps for calculations in warning charts. The steps are given in some detail because some modern investigators might be unfamiliar with using logarithms. Calculations are easily done with any computer spreadsheet; which will handle logarithms, antilogarithms, and calculate the mean and standard deviation. Indeed, the calculations are simple enough on a hand-held scientific calculator.

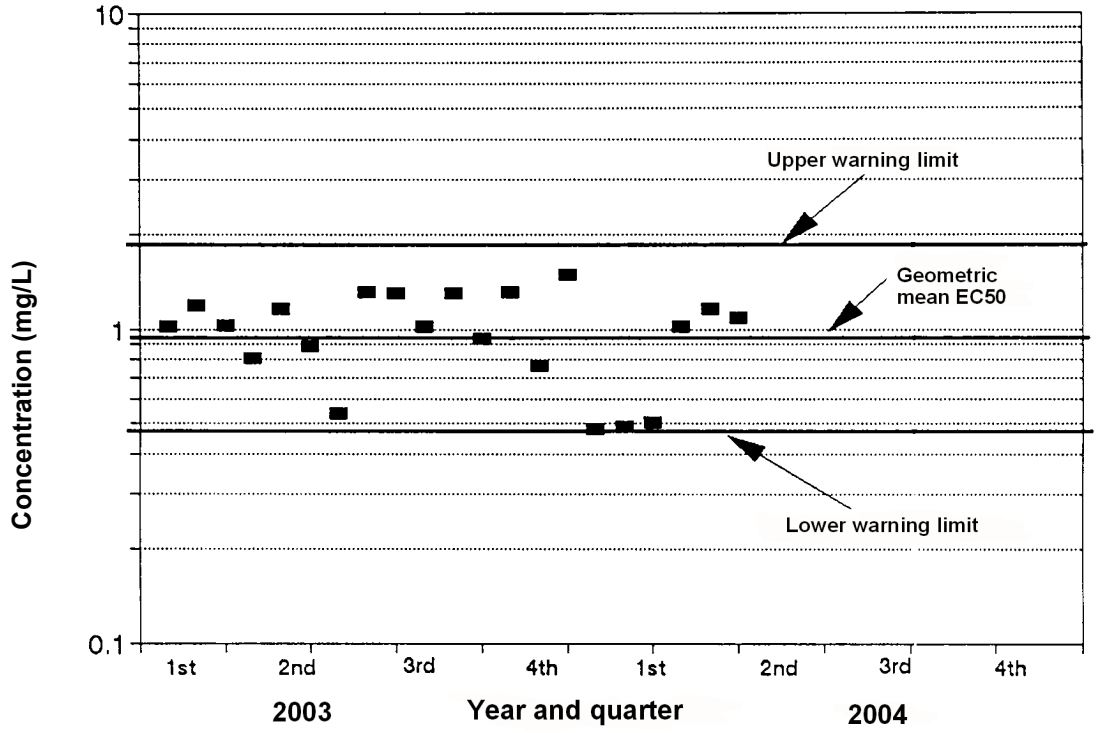


Figure F.1 Warning chart for tests with a reference toxicant. The chart/graph shows data from aquatic tests with a reference toxicant, from a Canadian laboratory.

The steps for calculating warning limits are as follows.

(1) Compile the historic data. These are the previous EC50s estimated for the reference toxicant, at the laboratory. Probably the EC50s would be recorded as arithmetic values, so each would be converted back to a logarithm. (Log to the base 10 is customary, although natural logarithms are equally good if they are used throughout.) For purposes of this example, only the first five EC50s from Figure F.1 are listed, and only a few digits are shown for the logarithms.

antilog EC50	1.02	1.19	1.03	0.81	1.16
EC50 as log	0.0086002...	0.075547...	0.012837...	-0.091515...	0.064458...

(2) The logarithms are averaged. This is simply the arithmetic *mean* of the logarithms.
Using all 21 EC50s of Figure F.1, the mean of logarithms is -0.027356... That mean remains as a logarithm for subsequent calculations, but is more easily comprehended when translated to 0.93895 mg/L (before rounding). The arithmetic value 0.94 is the *geometric mean EC50*. It is plotted as a line in Figure F.1.

(3) The standard deviation is calculated for the 21 logarithms of the EC50s.
This turns out to be 0.15288...

(4) The value of two standard deviations is twice the value in step (3).
2 x 0.15288... = 0.30576...
The antilog of this before rounding is 2.0219, but the antilog is of no particular use.

(5) The upper warning limit is calculated as the mean (step 2) plus 2SD (step 4).
-0.02736... + 0.30576... = 0.278404...
This can be converted to its antilogarithm of 1.9 mg/L, which is the upper warning limit and can be plotted on the control chart (see Figure F.1).

It is a mistake to do the calculations of steps (5) and (6) with arithmetic values; the wrong answers will be obtained. (However, see the following text for using arithmetic values with multiplication and division instead of addition and subtraction.)

(6) The lower warning limit is calculated as the mean (step 2) minus 2SD (step 4).
-0.02736... - 0.30576... = -0.33312...
Converted to its antilogarithm of 0.46 mg/L, this lower warning limit is plotted (Figure F.1.)

The warning limits are symmetrical about the mean in Figure F.1, because the vertical axis is a logarithmic scale. In the past, some investigators unfamiliar with concepts of logarithms have been distressed that warning limits calculated in the above fashion were not symmetrical when plotted on an arithmetic scale. That should not be a concern. **Correct limits will never be arithmetically symmetrical (they should not be), but they will be symmetrical on an appropriate logarithmic scale**.

There is another way of calculating the warning limits, if desired, using arithmetic values. Adding and subtracting logarithms is equivalent to multiplying and dividing their arithmetic equivalents.

- Thus, the upper confidence limit could be calculated as the geometric mean (step 2) multiplied by the antilogarithm of two standard deviations (step (4)):
0.938954 x 2.0219 = 1.9 mg/L, the same value obtained in step (5).

- The lower confidence limit could be estimated as the geometric mean divided by the antilogarithm of two standard deviations:
0.938954 / 2.0219 = 0.46 mg/L, once again the same value obtained in step (6).

There is also an alternative way of graphing the data. An arithmetic scale could be used for the vertical axis, and the logarithmic values could be plotted. Most investigators would probably consider this more cumbersome. Spreadsheets make it simple to plot the values on a graph with a logarithmic scale.

It is worth comparing the erroneous warning limits that would have been obtained if logarithms had not been used, i.e., if calculations had been based on the arithmetic values of the EC50s: 1.02, 1.19, 1.03, 0.81, etc.

- The average would have been calculated as 0.99 mg/L, somewhat higher than the proper value of 0.94 mg/L.

- The warning limits would have been 1.6 instead of 1.9 mg/L, and 0.39 instead of 0.46 mg/L. Thus the warning limits would have been appreciably lowered on Figure F.1. The range between limits would have been smaller, at 1.2 mg/L instead of 1.4 mg/L.

At first, it might seem anomalous that the erroneous mean is higher than the logarithmic one, while the erroneous warning limits are lower than the ones calculated logarithmically. This is a foreseeable part of the distortion. The erroneous (arithmetic) limits are equally spaced above and below their mean on an arithmetic scale. The confidence limits derived logarithmically are not equidistant from their mean on an arithmetic scale, but they are properly symmetrical as multiples of the arithmetic mean, differing from it by a factor of about 2.0.

"Reasonable" variation in EC50s. As mentioned in Section 2.8.1, Environment Canada has offered advice that variation in repeated tests of a reference toxicant would be considered reasonable if the coefficient of variation (CV) were 30%, and preferably 20%. This guideline was derived by calculations with arithmetic endpoints, which is subject to bias and undesirable. Therefore, the guideline was converted to a logarithmic basis, in an approximate way, by the process outlined below. Extra significant figures were carried in the calculations, and in the following text, three dots after a logarithmic value indicate the omission of numerals that would normally be carried for logarithms.

Several real and "dummy" sets of EC50s were compiled. Using the arithmetic values of the EC50s, coefficients of variation (CV) were calculated. One of the sets of EC50s was adjusted so that CV = 30.0% and another so that CV = 20.0%. Then for each set of EC50s, the standard deviation (SD) was calculated using the logarithmic values of the EC50s. The arithmetic CVs and the logarithmic SDs showed an approximate straight-line relationship when plotted. Logarithmic SDs were picked from the relationship, to correspond with the arithmetic CVs of 30% and 20%.

The SDs were 0.132... and 0.0338... and they represent a translation of Environment Canada's rule of thumb for "reasonable" and "preferred" variation in a set of results. The same values apply to any set of results, because they were derived from ratios on a logarithmic scale. The actual (calculated) SDs for any set of logarithmic EC50s may be compared with those guideline values.

An actual SD of 0.153 can be calculated for the data shown in Figure F.1, which is higher than the "reasonable" value of 0.132 estimated previously. It may be concluded that the data of Figure F.1 are somewhat more variable than the "reasonable" guideline published by Environment Canada.

(If an SD equal to the guideline value of 0.132 actually prevailed for a set of data which had the same mean as the data in Figure F1, the warning limits would be somewhat more narrow than those in the Figure. The limits would be the mean of the data, ± 2 SD. The logarithmic mean being -0.027356... (see above), the limits would be -0.027356 - (2 × 0.132) and -0.027356 + (2 × 0.132). The results would be -0.2914... and 0.2366..., which have antilogarithms of 0.51 and 1.7 mg/L. Those limits for hypothetical data are a little narrower than the actual warning limits shown in Figure F.1, of 0.46 and 1.9 mg/L.)

(If an SD equal to the "preferred" guideline SD of 0.0338 prevailed for a set of data with the same mean as the data in Figure F.1, the warning limits would be much narrower. Going through calculations parallel to those of the previous paragraph, the warning limits would be 0.80 and 1.1 mg/L.)

These rules of thumb for reasonable and preferable variation among repeated toxicity tests might be seen as being somewhat optimistic.

Tests for Single-concentration Results with No Replication

Testing of this nature is usually done under a regulatory program such as monitoring of a waste discharge. A firm pass/fail criterion is used to judge the test results; however, statistical testing might be required. Some of the statistical tests are discussed here, to supplement the information in Section 3.

G.1 Fisher's Exact Test

Fisher's Exact test can be used for testing a single sample, along with a control, and without replication. Often the observed effect is mortality, so the data are quantal. Fisher's Exact test, which is applied to quantal data only, is used to compare the results. This is a one-tailed test of statistical significance, because the investigator is interested in whether mortality is higher in the test sample, than in the control. Such testing might also be appropriate for mortality in *Ceriodaphnia* tests (EC, 1992a).

The procedure can be seen in the following example, which represents the numbers of individual organisms in a single-sample test with no replication.

	Dead	Alive	Total
Test	6	4	10
Control	1	9	10
Total	7	13	20

The numbers of live and dead organisms are tabulated as shown. The null hypothesis is that the proportion of dead organisms is not greater in the test than in the control.

Subtotals for the rows and columns are entered into the margins. The total number in the table is called n, in this case $n = 20$.

The smallest of the four marginal subtotals is selected, in this case 7, and is designated as m_1. Then in the other margin (the margin which does not contain m_1), the smallest marginal subtotal is selected and called m_2. In this case m_2 is 10; the lower 10 (for the control) was taken although the same final result will be obtained whichever 10 is selected. Now the number is selected in the body of the table, that contributes to both m_1 and m_2. This number is 1, and it can be called f.

The next step is a comparison of *"f"* with critical values contained in a rather complex table, given in some statistics texts such as Zar (1999; "Critical values for Fisher's exact test"). The table is entered at a certain point according to the selected level of significance (usually a probability value of 0.05), and also according to the values of n, m_1, and m_2. At that location in the table, there will be two pairs of critical values of f, and investigators should use the first pair, which is for a one-tailed test. (The second pair is for a two-tailed test which is not relevant here). If f is less then or equal to the first critical value, or is greater then or equal to the second critical value, then the null hypothesis is rejected, and mortality is concluded to be greater in the test than in the control.

In this case, critical values from the table are 1 and 6. The calculated f of 1 is equal to the first critical value; therefore, the null hypothesis is rejected, and the test sample showed significant increased mortality. (If the upper row of the table had been used to select 10 as m_2, then f would have been 6, equal to the second critical value, with the same rejection of the null hypothesis.)

The test mortality of 6 out of 10, shown in the table, happens to be the lowest that would be significant, for a small control effect of one out of 10. If the test had shown 5 dead out of 10, the null hypothesis would have been accepted. This is not entirely out of line with the conclusion of Zajdlik *et al.* (2001) that a "pass/fail decision is ambiguous when 4 to 7 fish die" out of 10. If there were no control effect, lesser mortalities than the 6 out of 10 tested, namely 5 out of 10 and 4 out of 10, would be significant. If the control effect were higher, say 2 out of 10, higher mortalities (7 or more out of 10) would be required for significance.

An alternative diagrammatic and tabular method is "Finney's Tables", outlined in Section G.2.

G.2 Comparison with "Finney's Tables"

Mortality in a single group can be compared to a control, using the diagrams in Figure G.1, or by tables in Finney *et al.* (1963) from which they were derived. The diagrams shown are designed for 3 to 10 individuals in each group. The diagrams are provided by Wardlaw (1985) in a statistics textbook that is very friendly to non-statisticians and they only work if the numbers of organisms are equal in the test group and the control.

Figure G.1 can be used to test the previous example given, with mortality of 6 out of 10 in the test group and 1 out of 10 in the control. The diagrams are set up for a one-tailed test of significance, so the null hypothesis is that mortality in the test group is not greater than in the control.

The lower right diagram is used for the 10 × 10 comparison. Only the "numerators" are used to enter the diagram, i.e., 6 for the test and 1 for the control. Enter the diagram in the column marked 6 (for the test mortality of 6), and at the row marked unity (for control mortality of 1). At the intersection, the square is stippled, indicating that the probability of this occurring due to chance alone is 0.05 or less. The null hypothesis is rejected at that level of probability, and it is concluded that the test mortality is greater than the control mortality. (Note that the conclusion is not that "the two groups are different", implying a two-tailed conclusion in which the test group might be either higher or lower than the control.)

The black areas of the diagram are for combinations in which the probability is 0.01 or less. The white areas indicate probabilities greater than 0.05, i.e., the test group would not be significantly higher than the control by the usual critical value of P.

For combinations other than those shown in Figure G.1, investigators could turn to the tables of Finney *et al.* (1963). The tables cover not only comparisons of equal numbers of organisms, but all possible combinations of unequal numbers up to 40 per group. For example, the tables allow a test mortality of 18 out of 32 to be compared with a control effect of 2 out of 20. Wardlaw (1985) also explains a tedious arithmetic method of making a comparison, which would quickly become unmanageable as numbers rose above 10 in the groups!

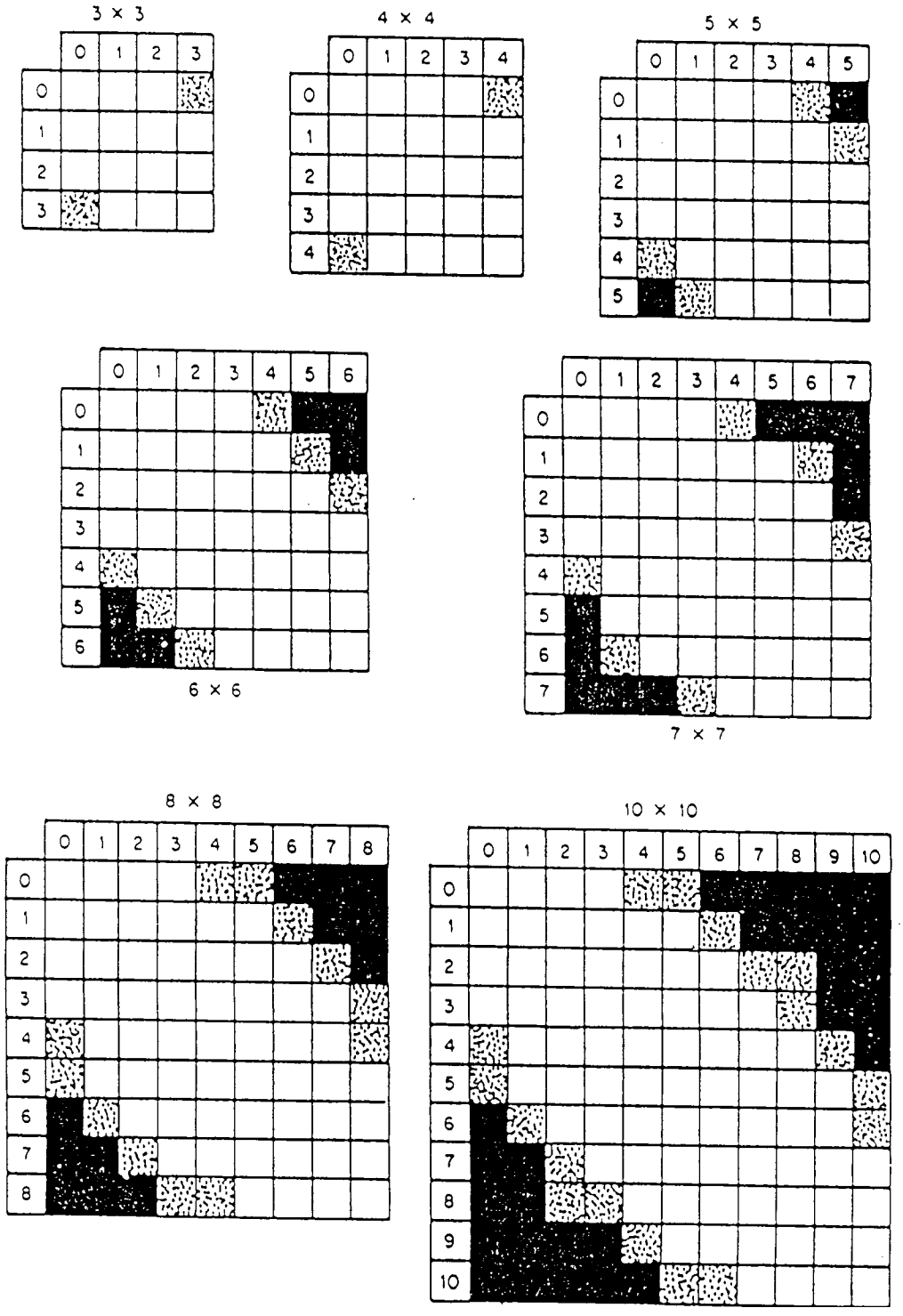


Figure G.1 Diagrams for comparing quantal effects in a test group and a control. The diagrams determine whether a test group shows a significantly greater effect than the control group. These diagrams are for equal numbers of experimental units (organisms) in test and control groups, from three in each (upper left diagram) to ten each (lower right). Black areas indicate $P \leq 1\%$, stippled blocks are $5\% \geq P > 1\%$, and white areas $P > 5\%$. After Wardlaw (1985), from tabulations of Finney *et al.* (1963).

G.3 Comparing Two Proportions with a Z-test

The method is given in general statistics texts, usually as "differences between proportions" or "comparisons of proportions" (e.g., Zar, 1999; Snedecor and Cochran, 1980). The method can be illustrated with the same data that used for the Fisher's Exact test.

	Dead	Alive	Total	Proportion dead
Test	6	4	$10 = n_T$	$0.6 = p_T$
Control	1	9	$10 = n_C$	$0.1 = p_C$
Total	7	13	20	
Proportion	$0.35 = p_{TC}$	$0.65 = q_{TC}$		

The statistic Z may be calculated by substituting in the following formula.

$$Z = \frac{p_T - p_C}{\sqrt{\frac{p_{TC}q_{TC}}{n_T} + \frac{p_{TC}q_{TC}}{n_C}}} = \frac{0.6 - 0.1}{\sqrt{\frac{0.35\times0.65}{10} + \frac{0.35\times0.65}{10}}} = \frac{0.5}{\sqrt{0.0455}} = 2.34 \qquad [\textit{Equation G.1}]$$

The critical value of Z for p = 0.05 and a one-tailed test is the same as the critical value of "t"for infinite degrees of freedom = 1.645. The calculated Z is greater than the critical value, therefore the null hypothesis is rejected, and mortality in the test chamber is greater than in the control.

Explanation of Probits and the Logarithmic-probability Transformation

H.1 Customary Transformations

Computer programs for probit regression use the log-probit transformation, which is seen in Figures 5, 8, and 9 in the main text. The transformation is intended to produce a straight line from what would otherwise be a skewed (see Glossary), cumulated normal curve (Figure H.1).

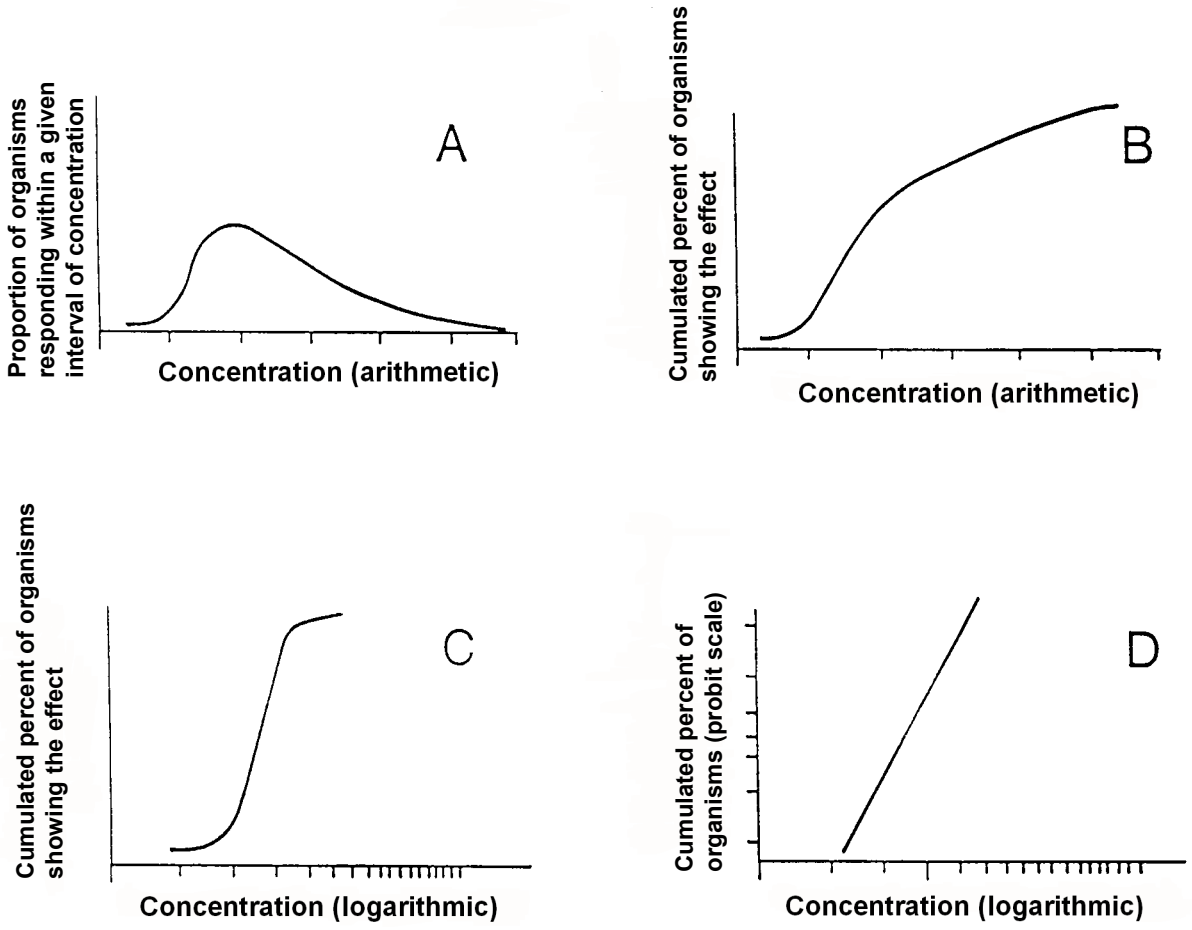


Figure H.1 Transformation of quantal data. Raw data from a test such as lethality to fish usually produce a skewed normal curve when plotted on arithmetic axes (Panel A). That distribution may be cumulated to produce a skewed sigmoid curve (Panel B), and logarithm of concentration removes the skew (Panel C). Applying a probability transformation to the percent effect (Panel D) straightens the line by vertically compressing the central portion and progressively extending the distal portions, which never reach 0% or 100% in this transformation.

If results of a quantal toxicity test were plotted on arithmetic paper, the result would almost always be a skewed normal curve. Panel A of Figure H.1 represents this as the proportions of total test organisms that showed the effect for each of a series of concentration-intervals. On the left of the curve, a few individuals are sensitive and show the effect at low concentrations. On the right, similar small numbers are very resistant, showing the effect only at very high concentrations. Most organisms are affected at the middle ranges of concentrations. If the numbers affected are cumulated, that yields a sigmoid or "S" curve, skewed to the right (Panel B).

Plotting logarithms of concentration usually eliminates the skew (Panel C). Using a probability transformation (= probit transformation) produces a straight line as shown in Panel D. The straight line allows easier techniques to be used for fitting the distribution of data, which was important during the development of new procedures, and in the past when calculations were done by hand or mechanical calculator. Today, complex calculations can be done on computer, so that the probit transformation could be omitted. Nevertheless, the older standard method with log-probit transformation continues to be a good model for hand-plotting a graph to check the pattern and the reasonableness of computer calculations.

H.2 Why Logarithms?

In a plot based on an arithmetic scale of concentrations, such as that for the raw data in Panel A of Figure H.1, the skew to the right is caused by the fact that a given arithmetic increase will represent successively diminishing proportions of higher concentrations.

A logarithmic scale adequately deals with this problem of changing proportions, since an increase by a given ratio in any arithmetic value (10 to 20, 100 to 200, or 1000 to 2000) results in the same numerical increase in a logarithm (Section 2.3). Or, on the logarithmic axis of a graph, a doubling of concentration occupies the same absolute distance, no matter where the starting point is located on the axis. This is true for logarithms to the base 10, and also for natural logarithms to the base "e". Base-10 logs are routinely used in environmental toxicology, and it is important not to mix types of logarithms in any given analysis.

H.3 What is a Probit?

Probits are equivalent to standard deviations from the standard normal distribution. Indeed they were originally called *Normal Equivalent Deviates* (*NED*; Gaddum, 1953), a name which had meaning to mathematicians, but is seldom seen today. In analysis of quantal toxicity data, probits substitute for the cumulated percent effect.

Probits are based on the usual distribution of frequencies in a normal curve: ± one standard deviation about the mean value includes about 68% of the observations; ± two standard deviations includes 95% of the observations; etc. If a cumulated normal curve is drawn (sigmoid), the theoretical relationship between cumulated percentages and standard deviations is still known. That relationship is used with probits.

A probit of value 1 (or one probit) corresponds to one standard deviation in the standard normal distribution (a normal distribution with mean = 0 and variance = 1.

Rather than the formal consideration, some simplistic diagrams can be used to illustrate the derivation of probits (Figure H.2). The panels are explained in the following steps.

(1) Start with a standard normal curve (Panel A of Figure H.2). Plus and minus one standard deviation from the mean includes 68% of the population (by definition of a normal curve). Plus and minus 2 σ includes 95% of the population, and + and - 3 σ includes 99.7%, etc.

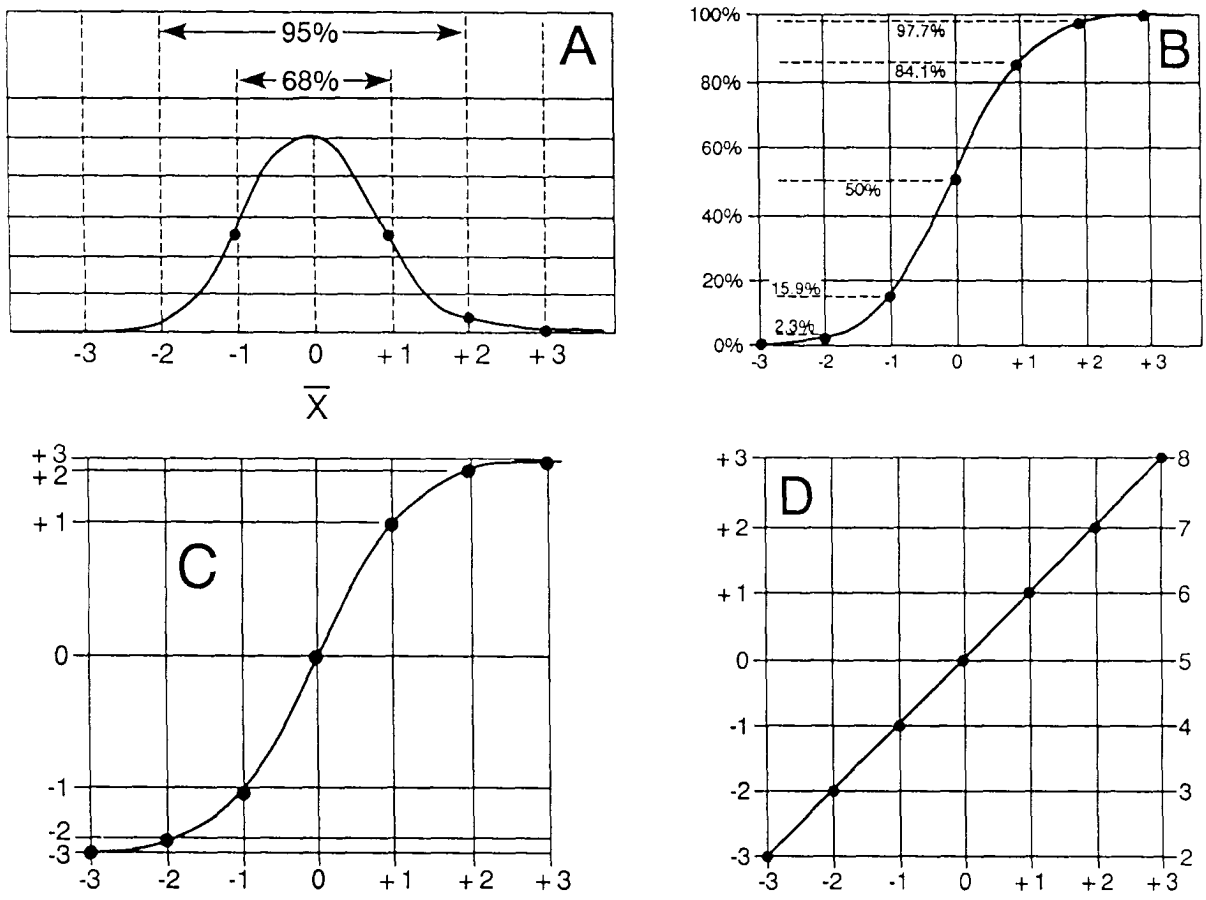


Figure H.2 The origin of probits. See text for explanation of the panels.

(2) Cumulate the curve. The percentages happen to work out as in Panel B of Figure H.2, shown on the dashed lines at various heights on the graph. This is a typical sigmoid curve.

(3) Now delete the percentage scale on the vertical axis, and number the dashed lines with the same numbers that are on the intercepts on the horizontal axis (Panel C). The numbers of the horizontal axis represent standard deviations.

(4) The vertical scale in Panel C is irregular with respect to the new numbering system. Use an arithmetic scale for the vertical axis with respect to the new numbers, running from -3 to +3 in the example of Panel D. The result is a straightening of the sigmoid curve. If the percentage scale were still present, it would be irregular, but the scale based on standard deviation is regular and the line is straightened.

This small exercise takes away the mystery. This is merely a method of fixing a cumulated normal curve, so that it looks straight. The units of the vertical axis have been transformed from percentages to equivalents of standard deviations, originally called *Normal Equivalent Deviates*, and now called *probits*.

One more modification has been customary, and investigators should be aware of it.

(5) The scale running from a negative to a positive value was awkward in the days of hand calculators. Therefore, each value had 5 added to it, so the usual working range became 2 to 8, as shown on the right side

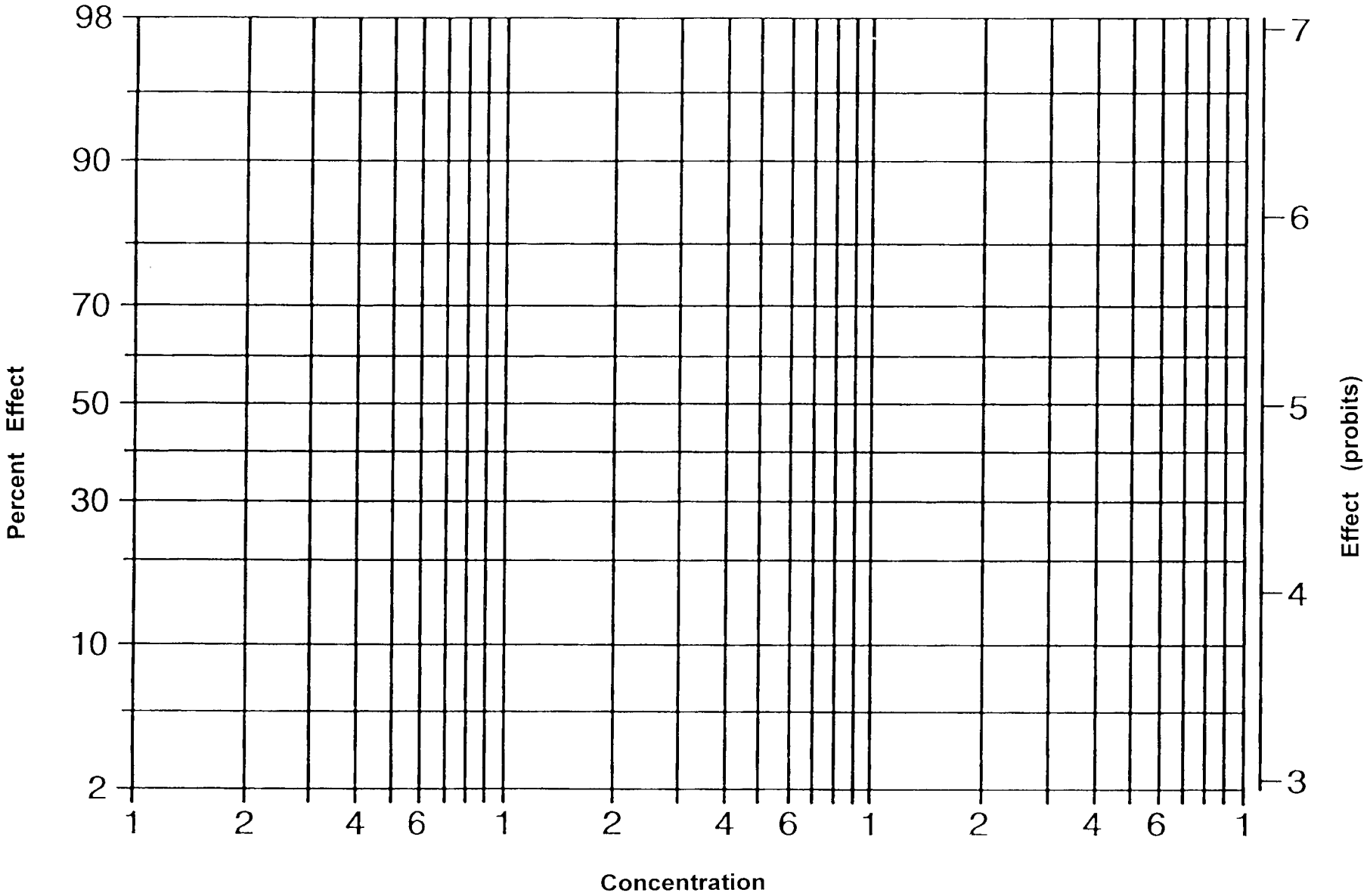
of Panel D. Thus, a probit of 5 became the midpoint. Strictly speaking, the definition of "probit" includes the added value of 5. Adding the 5 is no longer necessary for calculations by computer, but no harm is done if it is included.

Clearly there are relationships among probits, percentages, and standard deviations for a normal curve, so investigators can go from one to another if desired. The probit for any particular percentage can be found in published tables (Finney, 1971; Hubert, 1984; 1992), or obtain it from a normal probability calculator which is found in most statistical packages and spreadsheets. Computer programs for probit regression calculate the values.

Blank Logarithmic-Probability Paper (Log-probit Paper)

A sample of log-probit paper is provided on the next page. Photocopies of this could be used for analyses, if such paper is difficult to find. This paper is suitable for plotting the results of quantal toxicity tests. The effect is plotted on the vertical axis. Any quantal effect could be assigned, such as lethality, percent fertilization of salmonid eggs, or percentage of organisms showing lesions.

Various axes are available on commercial log-probit paper. In some, the probit scales run to very low and high values (e.g., 0.1% and 99.9%) which would be too extreme for most practical purposes.


98
90
70
50
30
10
2
Percent Effect
1
2
4
6
1
2
4
6
1
2
4
6
1
Concentration
7
6
5
4
3
Effect (probits)

Advantages and Explanation of Logits

Logistic methods are recommended over probits, for mathematical simplicity and other good reasons. However, both methods are good for analysis of quantal data, and the endpoints are usually very similar (Section 4.4).

Analysis of quantal data with logits is superior to using probits for several reasons.

- Numerically more stable than estimates with probits; failure is less likely (Hoekstra, 1989).
- The parameters produced by logistic regression make use of all the relevant information in a series of observations, which is not true in probit regression. Conversely, the parameters of a logistic regression have a direct meaning in reconstructing the original data.
- The parameters of the logistic model are widely used in the biomedical literature as measures of risk.
- Computer programs are somewhat easier to program for logistic regression models.
- Many more statistical packages are available, compared to probit regression.

The cumulated logit and normal distributions are similar (Figure J.1). Accordingly, the logit transformation can work satisfactorily on data that are normally distributed and suitable for probit regression. (Section 4.5.1 describes how the binomial effect in each individual test container became suitable for analysis using a normal or logistic distribution, when the cumulative distribution for all containers was considered.) Figure J.1 shows that the logistic curve has wider tails that are "heavier" than the normal curve. If an investigator is interested in the tails (say, <5% or >95%), the logit and probit endpoints would be appreciably different.

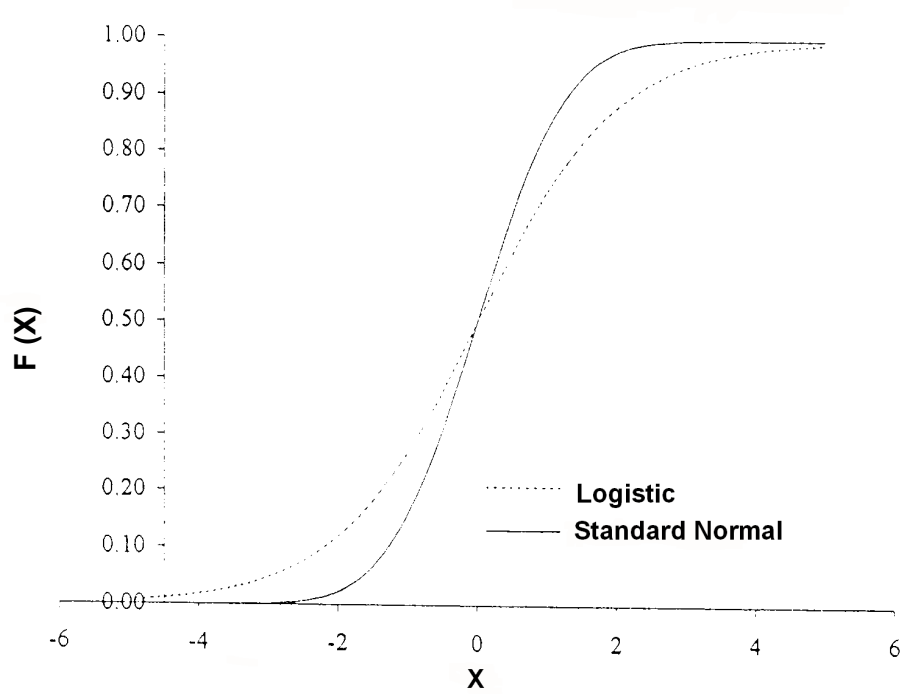


Figure J.1 Comparing the logistic and normal distributions. The distributions are cumulated, as done for the results of a quantal toxicity test.

The curves in Figure J.1 have been standardized about a mean of zero on the horizontal axis. For ease of understanding, the data may be treated as if they were weights of organisms, rather than as quantal data. Thus Figure J.1 represents the cumulated proportions of organisms with various weights. Standardization of the normal curve was done by subtracting the mean weight from each individual observation of weight, then dividing that by the standard deviation of the data-set. As a result, the horizontal axis is unitless, and has simply been labelled "x". For the normal curve, the values of "x" are standard deviations, the usual measurement of variability. To compare the logistic curve on an equivalent basis, the "scale" of the horizontal axis was set at unity [67]. For both curves, the vertical axis, F(x), describes the probability of obtaining a value less than "x"; that is, F(x) is a function that integrates the area under the curve to the corresponding point on the x-axis.

Some insight into the relationship of probits to a dose-effect curve, and how the curve is straightened, is provided Figure 9, which compares probits and logits.

A fairly simple transformation is used to obtain logits. For a given concentration, the proportion of organisms affected (p) is divided by (1 - p). The logarithm of the result is taken and that logarithm is the logit which can be used in fitting a regression and estimating an endpoint. The regression is linear and the equation becomes:

$$\text{logit } (p) = \alpha + \beta X$$

Thus, for quantal data such as a lethality test, the transformation has ended with a relationship that is similar to the familiar formula of a straight line *(simple linear regression)*: $Y = \alpha + \beta X$. This is, of course, the relationship between an effect (Y, the *dependent variable*) and X, the *independent variable* (the logarithm of concentration), further explained in Sections 6.5.1 and 6.5.2. This familiar formula, and the parallel equation for logits, represent a *regression* with only two parameters, **α** the intercept with the y-axis, and **ß** the slope of the line.

In both logistic and probit regression, for this quantal example, the parameters **α** and **ß** cannot easily be estimated because no equation can be written to solve for one parameter, that does not contain the other parameter. The solution is usually achieved by *iteration* (Section 4.5.3). For logistic regression, we may generalize that a computer program "guesses" at the value of the second parameter, solves the equation for the first parameter, then uses that estimate to solve for the second parameter. The process repeats, starting with the newly estimated value of the second parameter, until calculations are stopped by pre-determined criteria that indicate a satisfactory solution.

The regression, having been established for this quantal data, can be used to estimate the ECp and its confidence limits.

The logit transformation also provides a valuable model for sublethal quantitative data such as growth or reproduction. It has now been adopted by Environment Canada as an option for analyses of such sublethal tests. (see Section 6.5.4).

---

[67] For the normal curve, the "scale" in the sense of size, is in standard deviations. Thus, if dealing with weights of fish, the x-axis would be in terms of "standardized fish weights", made unitless. The scale of the logistic distribution is not the standard deviation, for rather complex statistical reasons. Setting the logistic scale at unity makes the comparison equitable. Statisticians would customarily refer to the units of the axis as "quantiles" and would label the axis with that descriptor.

The Spearman-Kärber Method

The Spearman-Kärber method (S-K) of estimating an EC50 has been widely used, particularly after it received support within the USEPA. The requirements of the method and its general approach are given in Section 4.5.6. Additional detail is given here on the inner workings of S-K, so that investigators can appreciate how the program processes the data.

Early methods documents published by Environment Canada did not recommend using S-K for estimating ECps because "divergent results might be obtained by operators who are unfamiliar with the implications of trimming off ends of the dose-response data" (EC, 1992b). In general, it was felt that S-K might manipulate test data in ways not understood or realized by the investigator, and the smoothing of irregular data might distort situations which deserved to be recognized as unusual. The famous statistician Finney also questioned S-K because it is "arithmetically possible to use it in situations where its validity is in grave doubt" (Finney, 1983, pers. comm., Dept. of Statistics, Univ. of Edinburgh, Edinburgh, Scotland). Indeed, anomalous results can be obtained for irregular data-sets (Section 4.4).

Environment Canada has recently recommended the untrimmed version of S-K to analyze tests showing one partial effect, which are not suitable for probit/logit regression (EC, 2001a, 2004a). A less restrictive approach is recommended herein. Use S-K for data-sets with only one partial effect, perform both untrimmed and minimally trimmed analyses, and select an ensuing endpoint that is reasonable, as judged from a plot of the raw data and the data themselves.

K.1 Simple Example Calculations

The basis of S-K is a process to calculate the mean of a probability distribution, in essence the mean of a frequency histogram. The mean is taken as the median, which is true for symmetrical distributions.

Table K.1 provides a *very* simplified example of a test with fish, to show how the method works. The data are shown for two concentrations, 10 and 20 mg/L, with zero effect out of 10 fish in the low concentration, and a complete effect at the higher concentration.

Using the S-K method, the EC50 is estimated to be 15 m/L. To give an anthropomorphic explanation, the lower concentration failed to kill any of the fish, but the higher concentration was sufficient to kill all of them. In essence, the method assumed that is there had been several intermediate concentrations, they would have killed the fish in regularly increasing proportions from zero to 100%. Therefore, the method assigned half of the mortality to the mid-point between the two concentrations actually used. Or, looking at this simplistic example in another way, the program assumed that the weakest fish would not be affected at 10 mg/L but would be at 11 mg/L, the next weakest not affected at 11 mg/L but at 12 mg/L, etc. Thus an effect on the fifth (median) fish would be predicted at 15 mg/L, and that was taken as the EC50.

The real procedure would use logarithms of concentration; arithmetic was used in this example for simplicity. It is customary to use natural logs ($\log_e$) with S-K, but the base of logarithms does not matter, as long as there is consistency.

Table K.1 A simplified example to demonstrate calculations of the Spearman-Kärber method. Arithmetic values of concentration are used to assist understanding.

(1)	Concentration	10 mg/L		20 mg/L
(2)	Mid-point of concentration range, $(C_1 + C_2) \div 2$		15 mg/L	
(3)	Proportion of fish affected	0.0		1.0
(4)	Proportion of fish dying in that interval of concentrations (1.0 - 0.0 = 1.0)		1.0	
(5)	Product of items (2) and (4)		15 mg/L	
(6)	EC50 = sum of all of items (5)		15 mg/L	

Usually there would be more concentrations, as in the more realistic example in Table K.2. This example has unusually large numbers of test organisms (fish in this case), and the proportions affected represent zero out of 40, 1/40, 1/40, 6/38, and 40/40. The example goes through exactly the same steps as in Table K.1, except that there are more concentrations and natural logs of concentration are used. The anthropomorphic explanation given above is also lost, as step (4) has four proportions of the total effect. Each of those proportions contributes to the final estimate of EC50, although in this case, most of the contribution comes from the right-hand proportion. It is important to carry many digits through the calculations.

Table K.2 A typical example of calculations by the Spearman-Kärber method.

(1) Concentration (mg/L)	15.54		20.47		27.92		35.98		55.52
ln_e conc.	2.7434		3.0190		3.3293		3.5830		4.0167
(2) Mid-point		2.8812		3.1742		3.4562		3.7999	
(3) Proportion affected	0.0		0.025		0.025		0.158		1.00
(4) Proportion in that interval		0.025		0.0		0.133		0.842	
(5) Product, (2) x (4)		0.07203		0.0		0.45967		3.19952	
(6) Total of items (5)									3.7312

The estimated EC50 is 3.7312 and its antilog is 41.7 mg/L. Confidence limits are calculated from variance and are 39.9 and 43.7 mg/L. Probit regression gives very similar results in this case.

K.2 Comments on Procedures

Smoothing of data is a manipulation used in S-K calculations, to obtain monotonic data. Smoothing can be necessary because the method requires that the effect at any given concentration *must* be greater than or equal to the effect at the next lower concentration. If not, the average effect at those two concentrations is taken, assigned to both concentrations, and used in calculations. This is called "adjusted proportion affected". In Table K.2, the two values of 0.025 listed for proportion affected had been previously adjusted from 0.05 in the second concentration, and zero in the third concentration.

Trimming the ends of the distribution is an option in computer programs for S-K (the "Trimmed Spearman-Kärber method"). The user can mathematically trim off 10%, 20%, or more, of the data at the ends of the cumulated effect-curve, where there could be irregularities, and work with the central portion. For the example in Table K.2, the "10%-trimmed estimate" of EC50 would be 42.8 instead of 41.7 mg/L; possibly a better estimate with narrower confidence limits. Some computer programs (TOXSTAT, CETIS) automatically select the minimum suitable trim, beyond the control of the investigator, which is considered satisfactory and is recommended here.

The validity of trimming has, however, been questioned. The original S-K method required effects of 0% and 100% at the ends of the distribution. If one or the other is missing, and both ends of the distribution are trimmed and discarded to get an even set of data, **the program then mathematically "expands" the distribution to 0% and 100%**, and proceeds to estimate the EC50. Trimming does not help if the irregularity is in the central part of the distribution. If such irregularities existed, it is the responsibility of the investigator to recognize them and deal with them appropriately.

K.3 Mathematical Formulae behind Spearman-Kärber Analysis

The formulae used in the Spearman-Kärber method are shown with two examples. Table K.3 shows calculations for Example A of Table 2 in the main text. Table K.4 shows another example in which smoothing took place. The comparison shows a major feature of Spearman-Kärber analyses, that the smoothing procedure tends to widen the confidence limits.

The log EC50 is estimated by:

$$u = \tfrac{1}{2} \sum_{i=1}^{k-1} (p_{i+1} - p_i)(x_i + x_{i+1}) \qquad \text{[Equation K.1]}$$

where:

p_i refers to the proportion dying (out of n_i organisms) at the ith concentration,
x_i refers to the i^{th} log concentration,
k is the number of concentrations,
p_1 = 0% mortality,
p_k = 100% mortality.

The variance of u is given by:

$$V(u) = \sum_{i=2}^{k-1} \frac{p_i(1 - p_i)(x_{i+1} + x_{i-1})^2}{4(n_i - 1)} \qquad \text{[Equation K.2]}$$

Confidence intervals are estimated as twice the standard deviation and so are constructed as EC50 ± 2 × standard deviation, which assumes that the estimated EC50 is distributed as a normal random variable.

Table K.3 Spearman-Kärber calculations for example A of Table 2.

Conc., (mg/L)	Log(conc.) (x_i)	i	n	No. Dead	Proportion dead (p_i)	Contribution to EC50 $(p_{i+1} - p_i)(x_i + x_{i+1})$	Contribution to variance (Equation K.2)
1.8	0.255273	1	10	0	0	0	
3.2	0.50515	2	10	2	0.2	0.07604	0.001080
5.6	0.748188	3	10	4	0.4	0.1253	0.001633
10	1	4	10	9	0.9	0.4370	0.0006430
18	1.255273	5	10	10	1	0.1128	
Sums :						log(EC50) = 0.7512	variance of log(EC50) = 0.003356

The approximate 95% confidence interval on the log(EC50) is ± 2 [square root of the variance of (log(EC50))]. That is 0.7512 ± 2 [square root (0.003356)] which estimates limits of 0.6353 and 0.8670. These values can be exponentiated to obtain EC50 = 5.64 with a 95% confidence interval of 4.32 to 7.36. These are essentially the values shown in Table 2.

If the effect is non-monotonic, it must be adjusted (smoothed) before using the Spearman-Kärber method. Adjacent effects are combined according to Equation K.3, which is tailored for the example in Table K.4.

$$p_{3.5} = \frac{e_3 + e_4}{n_3 + n_4} \qquad \text{[}\textit{Equation K.3}\text{]}$$

The data in Table K.4 can be described as a general case. The series of concentrations can be taken as c_1, c_2, c_3, c_4, and c_5. Taking "e" as the number affected and "n" as the number tested, the observed proportional effects are $p_1 = e_1/n_1$, $p_2 = e_2/n_2$, $p_3 = e_3/n_3$, $p_4 = e_4/n_4$, and $p_5 = e_5/n_5$. In this example, $p_3 > p_1$, p_2, and p_4, while $p_4 > p_1$ and p_2. It is required to combine p_3 and p_4 to obtain $p_{3.5}$, as in Equation K.3.

Since $p_2 < p_{3.5} < p_5$, calculations can proceed to estimate the endpoint. If monotonicity had not been obtained, smoothing would be repeated in the same fashion.

The EC50 and its 95% confidence interval are estimated as in Table K.3. The EC50 is 5.66 with a 95% confidence interval from 4.12 to 7.78.

The effects are similar in these last two examples, and the EC50s are approximately equal (5.64 and 5.66). The confidence interval is somewhat wider in the second case which used smoothing (4.12–7.78), than in the previous case (4.32–7.36). This is a typical consequence of the monotonization procedure.

Table K.4 Spearman-Kärber calculations for data that require smoothing.

Conc., (mg/L)	Log(conc.) (x_i)	i	n	No. dead	Proport'n dead (p_i)	Adjusted proport'n	Contribution to EC50 $(p_{i+1} - p_i)(x_i + x_{i+1})$	Contribution to variance (Equat. K2)
1.8	0.255273	1	10	0	0	0	0.114063	
3.2	0.50515	2	10	3	0.3	0.3	0.188001	0.001417
5.6	0.748188	3	10	7	0.7	0.6	0	0.001633
10	1	4	10	5	0.5	0.6	0.451055	0.001714
18	1.255273	5	10	10	1	1		
Sums :							log(EC50) = 0.753119	variance of log(EC50) = 0.004764

Background on Other Methods for Quantal Data

L.1 The Graphic Methods of Litchfield and Wilcoxon

This former "short-cut" method (Litchfield and Wilcoxon, 1949) is now a curiosity, but was frequently used until the 1960s, before easy access to electronic calculators or computers. The method is based on an eye-fitted line, but produces reasonable results. It estimates the EC50 and 95% confidence limits, the slope of the fitted line, and chi-square as an assessment of fit.

Outmoded as they are, the procedures are briefly described here in case an investigator can use them. Older work could be evaluated using this method, and it is useful for checking dubious output from computer programs. (In any case, the first part of the Litchfield/Wilcoxon method is an eye-fitted probit line, which is recommended in all analyses to determine an ECp, as a check on computer estimates.) It is instructional to try some of these analyses by hand, in particular, to see how the slope chosen for a probit line influences the width of the confidence limits about the EC50. Various lines can be tried for fit.

The procedures were designed to avoid the tedious hand calculations of probit regression. Slope of the eye-fitted line, and fit (chi-square) are calculated on the basis of deviations of the observed points from the line. The 95% confidence limits about the EC50 are determined by using nomograms; i.e., pre-calculated solutions to complex calculations, represented by three linear scales printed on a page in parallel fashion. A transparent ruler is placed appropriately on two of the linear scales representing known variables, and the answer (unknown variable) is read off the third scale where the ruler crosses it.

A modern description of the Litchfield-Wilcoxon method is provided by Newman (1995), who substitutes arithmetic calculations for the nomograms. The calculations are now easy enough on hand calculators, and the arithmetic procedures should be used to replace the nomograms of Litchfield and Wilcoxon.

L.2 Linear Interpolation

Section 4.5.9 points out that "linear interpolation" was designated as a particular technique for quantal data by the USEPA, but has no special use for Environment Canada tests. If a test shows no partial effects, investigators can use the binomial method which is exactly equivalent to linear interpolation. For other configurations of data, more suitable methods should be used as recommended in Section 4.3. The "linear interpolation" procedures of the USEPA are described here because they are frequently encountered in reports, and to explain why they are not required in Canada.

Early computer programs for linear interpolation were based on arithmetic values of concentration (Section 4.5.9), a deficiency that was remedied in more recent methods for testing dredged material, which use logarithms (USEPA and USACE, 1994).

The linear interpolation method merely interpolates between two points, and ignores other parts of the effect-distribution. If two successive test concentrations produced 0% and 100% effect, the calculations for linear interpolation could be done by Equation 3 (Section 4.5.7), the formula for geometric average.

A more generalized equation for linear interpolation handles results which show a partial effect at one or more concentrations. This might conceivably be useful in an unusual situation, although other methods are recommended here (Section 4.3). Equation L.1 is provided by USEPA and USACE (1994). Confidence limits cannot be obtained with this formula.

$$EC50 = antilog \frac{(50 - M_L)(\log C_U) + (M_U - 50)(\log C_L)}{M_U - M_L} \qquad [\ Equation\ L.1\ ]$$

where:

C_L = the arithmetic value of the concentration with effect nearest to and below 50% (the "lower" concentration)
C_U = the arithmetic value of the concentration with effect nearest to and above 50% (the "upper" concentration)
M_L = the percent effect at C_L
M_U = the percent effect at C_U

Nonlinear and Kernel Methods for Quantal Data

M.1 Nonlinear Regression

Kerr and Meador (1996) point to existing nonlinear techniques for estimating an ECp. Conventional analysis transforms to a linear relationship by means of probits (or logits), their example with a generalized linear model (GLIM) "... utilizes the inherent S-shaped feature of the toxicologic response ...". It is not clear whether the advantage of not needing a transformation would be outweighed by the disadvantage of needing more parameters in the equation fitting the relationship. Their model does, however, have the desirable feature of taking into account the sample size, and it also effectively uses 0% and 100% effects without any need for correction factors. The model can estimate ECp and its confidence limits for any value of p from low to high. This GLM uses an "... iteratively reweighted least-squares (IRLS) algorithm to find the parameter estimates that minimize the deviance." Kerr and Meador state that the analytical libraries SAS, Systat, and others have algorithms or specific programs for GLM, and can be used to estimate LCp. A certain degree of statistical knowledge is required to use the technique from those libraries.

Unfortunately, Kerr and Meador follow the same naive path as some others for their analysis, in abandoning the near-geometric distribution of test concentrations in the example data. The estimate of endpoint could well be accurate because the model can adapt to various curvatures and does not depend on a straight-line relationship. Still, abandoning the initial geometric/logarithmic assumption was not proper scientific procedure, and using that geometric base for concentration might have accomplished a fit that was more parsimonious in use of parameters, a distinct statistical advantage. This scientific fault could easily be corrected in the model to make it into a standard method.

M.2 Kernel Methods

A *kernel estimator* is a smoothing function that evens out a curve by applying an averaging procedure to points in the vicinity of any given point. The smoothing procedure is applied in turn, to each of the originally observed points, in order to produce a smoothed curve. An EC50 would be estimated at the fiftieth percentile from the smoothed curve, then related to the corresponding log concentration.

A weighting process is employed for the smoothing. For any given point, the nearest observations would be given the most weight, while those further away would be given less weight. There are several techniques for weighting points, and the most interesting ones are as follows.

- The *rectangular kernel*, in which points in the vicinity of the target point are given a weight of unity, and all other points are given a weight of zero (i.e., they do not contribute).

- The *triangular kernel*, in which observed points greater than a specified distance from the target point are given a weight of zero, while closer observations are assigned weights ranging from 0 to 1.

- The *Gaussian kernel*, in which weights follow the Gaussian or normal probability density function. This implies that all observations are included in the estimation of the target observation.

The analyst can choose a "span" or *bandwidth* for regulating the weights indicated previously. The choice of this span has a greater influence on the resulting smoothed curve, than does the choice of kernel function (Hastie and Tibshirani, 1990). These procedures, including optimal bandwidth selection, are discussed by Härdle (1991) and Scott (1992).

Kernel methods have advantages for use with toxicity tests, since they are nonparametric, and could be applied when there were no partial effects in the set of data. Potential methods have not yet been evaluated for the types of data that might come from Canadian environmental toxicity tests, but some assessments of relevance have been made (Kappenman, 1987). Müller and Schmidt (1988) ran evaluations of very large simulated sets of data (48 concentrations with 48 organisms per concentration). If the data were non-sigmoidal, the kernel analysis estimated an EC50 with a variance 40 to 70% smaller than would be obtained by probit regression, a very impressive feature. However, sigmoidal data would be more usual in test results, and for those, the variance was 20 to 30% larger than that for probit regression.

Point Estimates for Quantitative Data by Smoothing and Interpolation

N.1 Preparation for Analysis

The steps of the Smoothing and Interpolation method are given here in much more detail than in Section 6.4.2. Calculation of the ICp can be done by hand if desired (the explanation follows). The example uses weight of fish at the end of the test.

(1) Calculate the average weight of the fish held in each replicate of each concentration (including control). From the values for replicates, calculate the overall average weight for each concentration.

(2) Plot the average weights against a horizontal axis of the logarithm of concentration. This is the subjective check on the quality of data.

(3) Smooth the data if necessary. No smoothing is necessary if the overall average weight stays the same or decreases, at each step of increasing concentration, from the control to the highest concentration. If that is not true in any step, smoothing is necessary. The process must use the weighted average of the means (see following text).

- If the average weight at the lowest concentration is greater than in the control, take the mean of those two average weights, and use that mean for both the control and the lowest concentration.

- If the average weight at the second-lowest concentration is larger than at the average weight at the lowest concentration, take the mean of those two average weights, and use the mean for both the lowest and second-lowest concentrations. Repeat this for each ascending pair of concentrations until the highest concentration is reached.

- If the new average weights do not monotonically stay the same or decrease at each step of concentration, repeat the smoothing procedure for the appropriate pair(s) of concentrations, weighting each value that enters the averaging procedure, according to the number of original concentrations it represents [68].

- Repeat the preceding two steps until the set of results is monotonic.

- The new averages are used as input data for the analysis. In Equation N.1 (Section N.2), the symbol M indicates the new average weight, M_1 for weight in the control, and M_j for weight in a concentration to be named. All the original concentrations remain in the analysis, perhaps with a modified (smoothed) effect.

---

[68] The smoothing is done in a particular way. If the lowest concentration produced an average weight of, say, 14 units, higher than the control weight of 8 units, those weights would be averaged in the first round of smoothing. The result (11 units) represents the effect in both the control and the lowest concentration. The investigator would then leave the control and lowest concentration, and proceed to the two concentrations which were next above the lowest concentration, and so on, pairwise through the concentrations. The second cycle of smoothing would start again with the control; if the second lowest concentration had an average weight of 13, higher than the new value for the lowest concentration (and the control), then it would be averaged in with the 11, *and the result used for the control and the two lowest concentrations*. The new average would be weighted for the number of original observations involved, in this case the value 11 would have twice the weight of the 13. If each concentration had four observations (replicates), the calculation would be $[(8 \times 11) + (4 \times 13)]/12 = 11.7$. Alternatively, one could revert to the original observations and average them: $[(4 \times 8) + (4 \times 14) + (4 \times 13)]/12 = 11.7$. The value 11.7 now represents the effect in the control and each of the two lowest concentrations. Note also, the second lowest concentration was because it was higher than the *average* of the control and lowest concentration; it was not actually higher than the original weight for the lowest concentration. For that reason, if doing the process by hand, it would be best to do the smoothing for initial pairs of values, then repeat the cycle.

The smoothing operation can be a risky manipulation of the set of data, particularly if it is irregular or shows hormesis. Section 6.4.1 outlines potential problems. It is important to assess whether the calculated endpoint is reasonable, when compared with the original (raw) data.

N.2 Estimation of ICp

The method of estimation appears complicated when listed step-by-step, but it is merely a simple linear interpolation between the two concentrations which bracket the desired effect. The steps given below are followed by a formula for the same procedure. The analysis continues from step (3).

(4) Decide on the value of "p". Let us take 25, so that IC25 will represent the concentration associated with a 25% smaller weight than the control fish.

(5) Inspect the data to determine the two concentrations that bracket a 25% reduction in weight. From here on, the procedure uses only those two concentrations and the average weights associated with them.

(6) Calculate the weight that represents the endpoint. It is 75% of the weight of the control fish, i.e., calculate M_1 multiplied by 0.75.

(7) Subtract from the result of step (6), the weight (M_j) at the concentration immediately below the IC25. This will be a negative number.

(8) Take the average weight at the concentration immediately above the IC25, and subtract from it, the weight (M_j) at the concentration immediately below the IC25. This will normally be a negative number. Call this M_{Diff}

(9) Divide the result of step (7) by the result of step (8).

(10) Calculate the difference between the logarithms of the two concentrations immediately below (C_j) and immediately above the IC25. (It is important to subtract the logarithm of the lower concentration from that of the higher concentration). Call this C_{Diff}.

(11) Multiply the result of step (9) by the result of step (10). This is the upward movement of concentration to the IC25, from the concentration (C_j) immediately below it.

(12) Add the result of step (11) to the logarithmic concentration (C_j) immediately below the IC25. The result is the IC25 as a logarithm.

$$ICp = C_j + \frac{0.75M_1 - M_j}{M_{Diff}} \times C_{Diff} \qquad [\ Equation\ \ N.1\ ]$$

where:

C_j = The logarithm of the concentration immediately below the IC25.
C_{diff} = The difference between the logarithms of concentrations adjacent to the IC25, the higher one minus the lower one.
M_1 = The average effect (weight of fish) in the control.
M_j = The average effect in the concentration immediately below the IC25.
M_{Diff} = The difference between the average effect at the higher concentration minus the effect at the lower concentration (sign is important).

If there are no test concentrations both lower and higher than the ICp, then it cannot be estimated. It can only be said that the ICp is lower than the lowest concentration tested, or greater than the highest concentration, as the case may be.

N.3 Confidence Limits and the Computer Program ICPIN

A computer must be used for "bootstrapping", to obtain 95% confidence limits about the ICp. This involves calculating a series of ICps that might have been obtained, based on resamplings of the original observations (replicates). From the series of hypothetical ICps, it is possible to calculate reasonable confidence limits for the estimated ICp [69].

ICPIN runs on personal computers and is available within commercial packages; however, it is not proprietary, and copies are available from the USEPA [70]. ICPIN is easy to use, has clear instructions, and obvious handling of data [71]. An earlier version of the program, BOOTSTRP, should not be used.

ICPIN carries out all steps (1) to (11) listed in Section N.2, and raw observations are entered into the program. For correctness, **investigators must enter the logarithms of test concentrations**, rather than arithmetic concentrations as specified in the instructions of the program. At the end, the ICp and its confidence limits estimated by the computer can be converted from logarithmic to arithmetic values, for convenience in understanding them. Some commercial programs based on ICPIN offer a chance to transform the concentration (or "dose") to base-ten logarithms; however, **the investigator should make sure that the transformation is actually retained and used** in the calculations (Section 2.3.2 and Appendix N).

ICPIN handles up to 12 concentrations including the control, and up to 40 items per concentration. These "items" must be true replicates. For example, if weights were measured for 10 fish in a container at a given concentration, the weights would not be replicates; their total weight or average weight would be the measurement to enter into

---

[69] At least 240 new estimates of hypothetical ICps should be made. Each estimate arises from re-sampling the data from each concentration of the test, allowing any data-point to be selected more than once ("random resampling with replacement"). The computer program does the random sampling. For example, the effect-data entered into the program might be the total (or average) weight of fish for each of four replicate chambers at each concentration of a test. The computer would select four values to represent a concentration, from the four available weights of that concentration. It would select each of the four values from the same spectrum of four weights ("sampling with replacement"), so each sampling would likely include some weights twice or more, and fail to include one or more weights. A similar selection would be made for each concentration in the test, then a hypothetical ICp would be calculated. Then the computer would start over with another set of random selections from the same data, with another calculation of ICp, and so on.

Depending on the chance selections, there could be quite a variety of data-sets and ICps might be obtained. Greater variation in the original data results in a wider spread among the calculated ICps. The series of ≥240 hypothetical ICps will have its own distribution. The concentrations which mark off 2.5% of the hypothetical ICps, at either end of the distribution, are used to estimate the confidence limits of the ICp that was actually obtained in the experiment. The bootstrap technique was proposed by Efron (1982), and discussed by Marcus and Holtzman (1988).

If the limits came from only 80 bootstrap samplings, estimates might be unstable (USEPA, 1995). The earlier *BOOTSTRP* computer program tended towards optimistic narrow confidence limits, and that was noted in the minutes of the Canadian Statistics Advisory Group (Miller *et al.*, 1993). This tendency was especially evident when the number of replicates was small, such as two replicates per concentration.

[70] Source of the program is EMSL-Cincinnati, United States Environmental Protection Agency, 3411 Church Street, Cincinnati, Ohio, 45244, USA. In practice, since the program is not proprietary, many investigators have obtained copies of it from colleagues at some other laboratory, and as mentioned, it is a component of commercial toxicity programs.

[71] Ease of use is not necessarily true for commercial programs that incorporate ICPIN, as indicated in Appendix N.4.

ICPIN, as one replicate. It seems unlikely that 40 replicates would ever be available in tests done by the methods of Environment Canada. Equal numbers of items are not required in the various concentrations. The degree of impairment chosen as an endpoint can range in values of p from 1 to 99%.

The investigator must specify the number of resamplings in the bootstrap portion of the program. The number can range from 80 to 1000 in multiples of 40; at least 240 are usually recommended (Norberg-King, 1993), and there is no reason not to select a high number, say 800. If there are more than six data-entries (replicates) per concentration, the program provides "original" 95% confidence limits. If there are fewer than seven data-entries, the ICPIN program of 1993 (Version 2.0) provides original and "expanded" confidence limits, and the investigator should use the expanded values, which are an attempt to allow for over-optimistic estimates of limits by the bootstrapping.

The program output includes tables of data and preliminary calculations. The ICp from linear interpolation is provided and should be used. Confidence limits are printed, original or original and expanded as indicated previously. A "mean ICp" is also printed from the bootstrap sampling, with standard deviation; that is not the result of the toxicity test and must not be reported as such.

N.4 Commercial Programs with ICPIN

Commercial software packages contain versions of ICPIN, along with other programs for analysis of toxicity tests. Three common packages at the time of writing, are: TOXSTAT version 3.5 (1996), TOXCALC version 5.0 (1994), and CETIS (2001). These software packages have been used by investigators in Canada. Since the programs are changed from time to time, and new ones will become available, only general comments are warranted here.

The commercial programs usually follow the procedures of USEPA closely, and tend to produce information to satisfy requirements of the USEPA, sometimes on the government report sheets. The test methods and reporting do not necessarily satisfy requirements of Environment Canada. TOXCALC requires tedious entry of much accessory information that is not required.

Commercial programs tend to be written for application in current personal computers. The commercial packages were not as easy and obvious as ICPIN itself, for setup, entry of data, and analysis. Some old versions of the commercial programs were peculiar in their methods of data entry, or recalcitrant in their operation. Manuals failed to cover topics or failed to offer understandable words. The commercial programs did not offer free telephone access for help. Investigators should run example files if they are provided, to get acquainted with the requisite format.

As mentioned, investigators must ensure that log concentration is used in the analysis, which will probably require entering logarithms for most software. TOXSTAT 3.5 offered a transformation of concentration to logarithms, but in order to retain the logs in the analysis, the operator must choose that option and then command that instruction to "run", before proceeding to the analysis.

Estimating ICps Using Linear and Nonlinear Regression

Section 6.5.8 gives the general procedures for Environment Canada's standard method of regression for quantitative data from toxicity tests. This appendix offers step-by-step generic instructions for carrying out an analysis. The statistical methods are identical to that included as standard procedure in Environment Canada's recent methods for soil toxicity (2004a–c).

O.1 Introduction

This appendix provides instruction for the use of linear and nonlinear regression analyses to derive ICPs, based on the concentration-response relationships for quantitative data. It represents an adaptation and modification of the approach described by Stephenson *et al.* (2000). Instructions here are for using Version 11.0 of SYSTAT[72]; however, any suitable computer software may be used. These regression techniques are most appropriately applied to continuous data from test designs with 10 or more concentrations or treatment levels including the control treatment(s). The test design for measuring the effects of prolonged exposure on the earthworm *Eisenia andrei*, Collembola (springtails, e.g., *Folsomia candida* or *Onychiurus folsomi*), and plant growth, is summarized in Table O.1.

An overview of the general process used to assess the suitability of a set of data for these regressions is presented in Figure 16.

Before data are analyzed, the reader should refer to appropriate sections within this statistical guidance document, as well as relevant sections on test design and regression analyses in the methods specific to earthworms, plants, and Collembola (EC, 2004a–c). Some of the related guidance from those documents has been provided in this appendix.

O.2 Linear and Nonlinear Regression Analyses

O.2.1 Creating Tables of Data

The statistical analysis must use the logarithms of concentrations ($\log_{10}$ or $\log_e$). If the concentrations fall below unity (1.0, e.g., 0.25), then the units of concentration can be changed (e.g., from mg/kg to µg/kg) using a multiplication factor (1000 in this example); the modified concentrations are then expressed as logarithms. Logarithmic values can be recorded in the original electronic spreadsheet, or the change can be done when the original data are transferred to the SYSTAT data file. The ICps and confidence limits should be transformed to arithmetic values for reporting, for ease of comprehension.

(1) Open the appropriate file containing the data-set in an electronic spreadsheet.

(2) Open the SYSTAT program. In the main screen, go to **File**, **New**, and then **Data**. This will open up an empty data-table. The user must first insert the variable names into the column heading by double-clicking on a variable name, which opens the '**Variable Properties**' window. Insert an appropriate name for the variable of interest within the '**Variable name**' box, and select the variable type; additional comments can be inserted within the '**Comments:**' box. For example, the following variable names might be used:

[72] The latest version of SYSTAT™ (e.g., Version 11.0) can be purchased from SYSTAT Software, Inc., 501 Canal Boulevard, Suite C, Point Richmond, Calif. 94804-2028, USA, phone 800 797-7401; web site www.systat.com/products/Systat/.

Table O.1 Summary of test design for Environment Canada's biological methods for testing toxicity of soil for plant growth, or reproduction of earthworms and collembolans.

Variable	Earthworm	Plant	Collembola	
Species	*Eisenia andrei*; sexually mature adults with clitellum and individual wet weights ranging from 250 to 600 mg	Various	*Folsomia candida*; age synchronized; 10–12 days after eclosion	*Onychiurus folsomia*; adults >2 mm body length; not age-synchronized; 5 males and 5 females
Test duration	56 days = 8 weeks	14 or 21 days; species-dependent	28 days	35 days
Number of replicates	10 replicates per treatment	6 per control treatment; 4 for each lower concentration; 3 for middle to highest concentrations	≥3 replicates per treatment; ≥5 replicates per control treatment	≥10 replicates per treatment, including the control treatment
Number of treatments	Negative control soil and ≥7 concentrations; ≥10 concentrations plus negative control strongly recommended	Negative control soil and ≥9 test concentrations as a minimum	Negative control soil and ≥7 concentrations; ≥10 concentrations plus negative control strongly recommended	Negative control soil and ≥7 concentrations; ≥10 concentrations plus negative control strongly recommended
Statistical endpoints	**Quantal** Methods in this appendix are not appropriate. Use quantal procedures if there is a suitable concentration-effect relationship.			
	• Mean percent adult survival in each treatment on Day 28 • Calculate 28-d LC50 (quantal procedures)	• Mean percent emergence in each treatment • Calculate 14- or 21-day EC50 by quantal procedures	• Mean percent adult survival in each treatment on Day 28 • Calculate 28-d LC50 (quantal procedures)	• Mean percent adult survival in each treatment on Day 35 • Calculate 35-d LC50 (quantal procedures
	Quantitative Estimate the ICp (e.g., IC50 and/or IC25)			
	• Mean number and dry mass of live juveniles in each treatment, on Day 56 • ICp for dry mass and number of live juveniles	• Mean shoot and root length and dry mass in each treatment, on Day 14 or 21 • ICp for the mean shoot and root length and dry mass	• Mean number of live juveniles in each treatment, on Day 28 • ICp for number of live juveniles produced	• Mean number of live juveniles in each treatment, on Day 35 • ICp for number of live juveniles produced

conc = concentration or treatment level
logconc = $\log_{10}$ value of the concentration or treatment level
rep = replicate within a treatment level
juveniles = number of juveniles produced
jdrywt = dry mass of juveniles produced
mnlengths = mean length of plant shoots
mnlengthr = mean length of plant roots
drywts = dry mass of plant shoots
drywtr = dry mass of plant roots

(3) Transfer the data by copying and pasting each column from the electronic spreadsheet containing the concentrations, the replicates, and associated mean values, to the SYSTAT data-table.

(4) Save the data by going to **File**, then **Save As**; a **'Save As**' window will appear. Use appropriate coding to save the data-file. Select **Save** when the file name has been entered.

(5) Record the file name of the SYSTAT data-file in the electronic spreadsheet containing the original data.

(6) If the data (i.e., the test concentrations) need to be changed to logarithms, select **Data**, **Transform**, and then **Let...**. Once in the **Let...** function, select the column heading containing the appropriate header for the format desired (e.g., logconc), and then select **Variable** within the '**Add to**' box to insert the variable into the '**Variable:**' box. Select the appropriate code (e.g., L10 for $\log_{10}$ or LOG for the natural logarithm) in the '**Functions:**' box (the '**Function Type:**' box should be **Mathematical**), and then select **Add** to insert the function into the '**Expression:**' box. Select the column heading containing the arithmetic version of the data (i.e., 'conc' for concentration or treatment level), followed by **Expression** within the '**Add to**' box to insert the variable into the '**Expression:**' box. If a multiplication factor is required to adjust the concentration before the change to logarithms, this step can be completed within the '**Expression:**' box (e.g., L10[conc*1000]). Select **OK** when all of the desired actions are complete. The logarithmic data will appear in the appropriate column. *Save the data* (i.e., select **File**, followed by **Save**).

The $\log_{10}$ of the negative control treatment cannot be provided, because the $\log_{10}$ of zero is undefined. Therefore, assign the control treatment level a very small number (e.g., 0.001) known or assumed to be a no-effect level. This will allow this treatment to be included in the analysis and will differentiate it from the other logarithmic treatment levels.

(7) From the data-table, calculate and record the mean of the negative controls for the variable under study. Each endpoint is analyzed independently. The mean value of these control data will be required when estimating the parameters of the model. In addition, determine the maximum value within the data-set for that particular variable and round it upwards to the nearest whole number. This number is used as the maximum value of the y-axis (i.e., '*ymax*') when creating a graph of the regressed data.

O.2.2 Creating a Scatter Plot or Line Graph

The scatter plots and line graphs provide an indication of the shape of the concentration-response curve for the data-set. The shape of the concentration-response curve can then be compared to each model (Figure O.1) so that the most appropriate model(s) for the data can be selected. Each of the selected models should be used to analyze the data. Each model should be reviewed subsequent to the analysis. Select the model that demonstrates the best fit.

(1) Select **Graph**, **Summary Charts**, and then **Line...**. Select the independent variable (e.g., logconc), followed by **Add** to insert the variable into the '**X-variable(s):**' box. Select the dependent variable under examination, followed by **Add** to insert the variable into the '**Y-variable(s):**' box. Select **OK**. A graph will be displayed within the '**Output Pane**' of the main SYSTAT screen containing the mean values for every treatment level. To view a larger version of the graph, simply select the '**Graph Editor**' tab located below the central window. A scatter plot of the data can also be viewed by selecting **Graph**, **Plots**, and then **Scatterplot...** and then following the same instructions for inserting the x- and y-variables. The graphs will provide an indication of the general concentration-response trend, allowing the potential model(s) of best fit to be chosen. They will also indicate the approximate ICp of interest.

The main SYSTAT screen is divided into three parts. The left-hand side ('**Output Organizer**' tab) provides a list of all of the functions completed (e.g., graphs) – each function can be viewed by selecting the desired icon. The right-hand side forms the central window in which the general output of all completed functions can be viewed (e.g., regression, graphs). The tabs below this central window allow the user to toggle between the data-file (use

tab '**Data Editor**'), individual graphs ('**Graph Editor**') and the output ('**Output Pane**'). The various graphs can be viewed individually within '**Graph Editor**' by selecting the graph of interest within the left-hand side of the screen ('**Output Organizer**' tab). The bottom portion of the screen displays the command codes used to derive the functions (regression and graphing codes). The '**Log**' tab within this command screen displays a history of all of the functions completed.

(2) Visually estimate and record an estimate of the ICp (e.g., IC50) for the data-set. For example, for an IC50, divide the average of the control measurements by two, and find this value on the y-axis. Project a horizontal line from the y-axis until it intercepts the data-points. Extend a vertical line downward to the x-axis and record this concentration as an approximate estimate of the IC50.

(3) Using the scatter plots or line graphs, select the potential model(s) that will best describe the concentration-response trend (refer to Figure O.1 for an example of each model).

O.2.3 Estimating the Parameters of the Model

(1) Select **File**, **Open**, and then **Command**.

(2) Open (or create) the file containing the command codes for the particular model chosen from Section O.2.2 (i.e., select the appropriate file, followed by **Open**):

```
nonline.syc   =  exponential model
nonling.syc   =  gompertz model
nonlinh.syc   =  logistic with hormesis model
linear.syc    =  linear model
nonlinl.syc   =  logistic model
```

The file will provide the command codes for the selected model within the appropriate tab of the command editor box at the bottom of the main screen. All of the command codes for deriving IC50s and IC25s are provided in Table O.2; however, the equations can be formatted to derive any ICp. For example, the command codes for the logistic model to derive an IC50 would be:

```
nonlin
print          =  long
model drywts   =  t/(1+(0.25/0.75)*(logconc/x)^b)
save resid1/ resid
estimate/ start  =  85, 0.6, 2 iter  =  200
use resid1
pplot residual
plot residual*logconc
plot residual*estimate
```

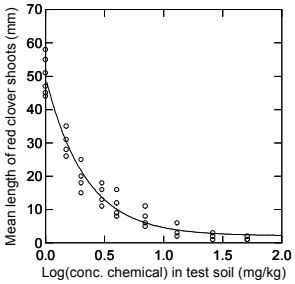
(3) For the column in the data-table which contains the variable to be analyzed, type the header within the line entitled 'model y=' (where 'y' is the dependent variable, e.g., jdrywt).

(4) The fourth line of the text should read 'save resida/ resid', where 'a' indicates a number to which the residual file is assigned. Substitute this same number into the sixth line ('use resida') so that the same file will be used to generate a normal probability plot and graphs of the residuals. The command lines that follow provide instruction for generating a probability plot ('pplot residual'), for generating a graph of residuals against the concentration or treatment level ('plot residual*logconc'), and for a graph of the residuals against the predicted and fitted values ('plot residual*estimate'). These graphs help to assess the assumptions of normality (e.g., probability plot) and homogeneity of the residuals (e.g., graphs of the residuals) when evaluating for the model of best fit (Section O.2.4).

Exponential Model

IC50: mnlengths = a*exp(log((a-a*0.5-b*0.5)/a)*(logconc/x))+b
IC25: mnlengths = a*exp(log((a-a*0.25-b*0.75)/a)*(logconc/x))+b

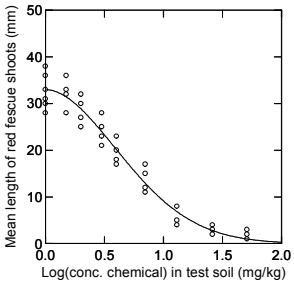
Where:
a = the y-intercept (the control response)
x = ICp for the data set
logconc = the logarithmic value of the exposure concentration
b = a scale parameter (estimated between 1 and 4)



Gompertz Model

IC50: mnlengths = g*exp((log(0.5))*(logconc/x)^b)
IC25: mnlengths = g*exp((log(0.75))*(logconc/x)^b)

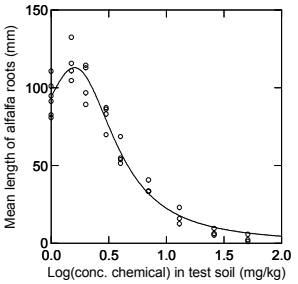
Where:
g = the y-intercept (the control response)
x = ICp for the data set
logconc = the logarithmic value of the exposure concentration
b = a scale parameter (estimated between 1 and 4)



Hormesis Model

IC50: mnlengthr = (t*(1+h*logconc))/(1+((0.5+h*logconc)/0.5)*(logconc/x)^b)
IC25: mnlengthr = (t*(1+h*logconc))/(1+((0.25+h*logconc)/0.75)*(logconc/x)^b)

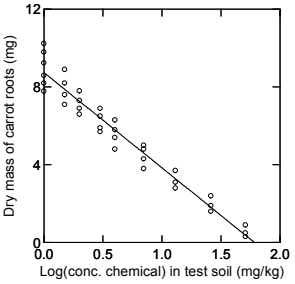
Where:
t = the y-intercept (the control response)
h = the hormetic effect (estimated between 0.1 and 1)
x = ICp for the data set
logconc = the logarithmic value of the exposure concentration
b = a scale parameter (estimated between 1 and 4)



Linear Model

IC50: drywtr = ((-b*0.5)/x)*logconc+b
IC25: drywtr = ((-b*0.25)/x)*logconc+b

Where:
b = the y-intercept (the control response)
x = ICp for the data set
logconc = the logarithmic value of the exposure concentration



Logistic Model

IC50: drywts = t/(1+(logconc/x)^b)
IC25: drywts = t/(1+(0.25/0.75)*(logconc/x)^b)

Where:
t = the y-intercept (the control response)
x = ICp for the data set
logconc = the logarithmic value of the exposure concentration
b = a scale parameter (estimated between 1 and 4)

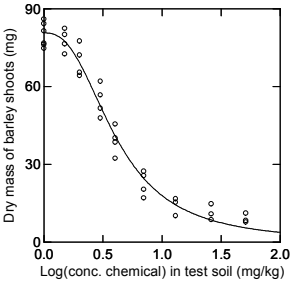


Figure O.1 Equations from SYSTAT Version 11.0 for linear and nonlinear regression models and examples graphs for each model

(5) Substitute the mean of the controls and the estimated ICp in the fifth line entitled 'estimate/ start=' (refer to Table O.2 for details on the substitution for each model). These values were initially derived from examining the scatter plot or line graph. The model, once it converges, will provide a set of parameters from which the ICp and its 95% confidence limits are reported (i.e., parameter 'x'). It is essential to provide accurate estimates for each parameter before running the model, or the iterative procedure might not converge. The estimate of the scale parameter (Table O.2) is usually between one and four. The number of iterations can be changed, but for this example it was set at 200 ('iter = 200'). Typically, 200 iterations are sufficient for a model to converge; if more than that are required, it is likely that the most appropriate model is not being used.

(6) Select **File**, and then **Submit Window** to run the commands; alternatively, right-click the mouse and select **Submit Window**. This will generate a printout of the iterations, the estimated parameters, and a list of the actual data-points with the corresponding predicted values and residuals. A preliminary graph of the estimated regression line will also be presented. This preliminary graph should be deleted by selecting the graph in the left-hand window within the main screen. A normal probability plot and graphs of the residuals will also be presented.

O.2.4 Examining the Residuals and Test Assumptions

An examination of the residuals for each tested model helps to determine whether the assumptions of normality and homoscedasticity have been met. If any of the assumptions cannot be met, regardless of the model examined, a statistician should be consulted for guidance on using additional models, or the data should be re-analyzed using the less desirable method of linear interpolation (ICPIN; Section 6.4 and Appendix N).

O.2.4.1 Assumptions of Normality

Normality should be assessed using *Shapiro-Wilk's test* as described in Section O.2.4.3 (see also Appendix P, Sections P.2.1 and P.2.2). The normal probability plot, displayed in the '**Output Pane**', can also be used to evaluate whether the assumption of normality is met. The residuals should form a fairly straight line diagonally across the graph; the presence of a curved line represents deviation from normality. The normal probability plot should not, however, be used as a stand-alone test for normality, because a decision on the degree of curvature would depend on subjective judgement of the user. If the data are not normally distributed, then the user should try another model, consult a statistician, or the data should be analysed using the less desirable linear interpolation method.

O.2.4.2 Homogeneity of Residuals

Homoscedasticity (or homogeneity) of the residuals should be assessed using *Levene's test* following the instructions in Section O.2.4.3 (see also Appendix P, Section P.2.3), and by examining the graphs of residuals. Homogeneity of the residuals is characterized by an equal distribution of the variance of the residuals, for all values of the independent variable (Figure O.2A). A significant result for Levene's test indicates that the data are heteroscedastic, and the graphs of the residuals should then be examined. If there is a significant change in the variance and the graphs of the residuals produce a distinct fan or 'V' pattern, then the data analysis should be repeated using weighted regression. (Refer to Figure O.2B for a plot of the '*residual*estmate*'; a corresponding 'V' pattern in the opposite direction also occurs in the plot of the '*residual*logconc*'.) Alternatively, a divergent pattern suggestive of a systematic lack of fit (Figure O.2C) will indicate that an inappropriate or incorrect model was selected.

O.2.4.3 Assessing Normality and Homogeneity of Residuals

SYSTAT Version 11.0 can perform both Shapiro-Wilk's and Levene's tests. Levene's test can only be performed by conducting an ANOVA on the absolute values of the residuals derived in Section O.2.3.

Table O.2 Command codes in SYSTAT for linear and nonlinear regression models

Model	Command Codes	
Exponential	nonlin print = long 'a') model mnlengths = a*exp(log((a-a*0.25-b*0.75)/a)*(logconc/x))+b save resid1/ resid estimate/ start = 25[a], 1[b], 0.3[c] iter = 200 use resid1 pplot residual plot residual*logconc plot residual*estimate	where: [a]Represents the estimate of the y-intercept (i.e., 'a') (the control response) [b]Represents the scale parameter (i.e., 'b') (estimated between 1 and 4) [c]Represents the estimate of the ICp for the data set (i.e., 'x')
Gompertz	nonlin print = long model mnlengths = g*exp((log(0.75))*(logconc/x)^b) save resid2/ resid estimate/ start = 16[a], 0.8[b], 1[c] iter = 200 use resid2 pplot residual plot residual*logconc plot residual*estimate	where: [a]Represents the estimate of the y-intercept (i.e., 'g') (the control response) [b]Represents the estimate of the ICp for the data set (i.e., 'x') [c]Represents the scale parameter (i.e., 'b') (estimated between 1 and 4)
Hormesis	nonlin print = long model mnlengthr = (t*(1+h*logconc))/(1+((0.25+h*logconc)/ 0.75)*(logconc/x)^b) save resid3/ resid estimate/start = 48[a], 0.1[b], 0.7[c], 1[d] iter = 200 use resid3 pplot residual plot residual*logconc plot residual*estimate	where: [a] Represents the estimate of the y-intercept (i.e., 't') (the control response) [b]Represents the hormetic effect (i.e., 'h') (estimated between 0.1 and 1) [c]Represents the estimate of the ICp for the data set (i.e., 'x') [d]Represents the scale parameter (i.e., 'b') (estimated between 1 and 4)
Linear	nonlin print = long model drywtr = ((-b*0.25)/x)*logconc+b save resid4/ resid estimate/start = 5[a], 0.7[b] iter = 200 use resid4 pplot residual plot residual*logconc plot residual*estimate	where: [a]Represents the estimate of the y-intercept (i.e., 'b') (the control response) [b]Represents the estimate of the ICp for the data set (i.e., 'x')
Logistic	nonlin print = long model drywts = t/(1+(0.25/0.75)*(logconc/x)^b) save resid5/resid estimate/start = 85[a], 0.6[b], 2[c] iter = 200 use resid5 pplot residual plot residual*logconc plot residual*estimate	where: [a]Represents the estimate of the y-intercept (i.e., 't') (the control response) [b]Represents the estimate of the ICp for the data set (i.e., 'x') [c]Represents the scale parameter (i.e., 'b') (estimated between 1 and 4)

(1) Select **File**, **Open**, and then **Data** to open the data-file containing the residuals created in Section O.2.3 (e.g., resid1.syd).

(2) Insert a new variable name into an empty column by double-clicking on the variable name, which opens the '**Variable Properties**' window. In this window, insert into the '**Variable name:**' box, an appropriate

name for the transformed residuals (e.g., absresiduals). Transform the residuals by selecting **Data**, **Transform**, and then **Let...**. Once in the **Let...** function, select the column heading containing the appropriate header for the transformed data (e.g., absresiduals), and then select **Variable** within the '**Add to**' box to insert the variable into the '**Variable:**' box. Select the appropriate transformation (e.g., ABS for the transformation of data into its absolute form) in the '**Functions:**' box (the '**Function Type:**' box should be **Mathematical**), and then select **Add** to insert the function into the '**Expression:**' box. Select the column heading containing the original untransformed data (i.e., residuals), followed by **Expression** within the '**Add to**' box to insert the variable into the '**Expression:**' box. Select **OK**, and then the transformed data will appear in the appropriate column. Save the data.

(3) To perform Shapiro-Wilk's test, select **Analysis**, **Descriptive Statistics**, and then **Basic Statistics...**. A '**Column Statistics**' window will appear. Select the residuals from the '**Available variable(s):**' box, followed by **Add** to insert this variable into the '**Selected variable(s):**' box. Within the '**Options**' box, select the **Shapiro-Wilk normality test**, followed by **OK**. A small table will appear within the SYSTAT Output Organizer window, where the Shapiro-Wilk critical value (i.e., 'SW Statistic') and probability value (i.e., SW P-Value') will be displayed. A probability value greater than the usual criterion of $p > 0.05$ indicates that the data are normally distributed.

(4) To perform Levene's test, select **Analysis**, **Analysis of Variance (ANOVA)**, and then **Estimate Model...**, an '**Analysis of Variance: Estimate Model**' window will appear.

(5) Select the variable within which the data are to be grouped (e.g., logconc), and place this variable into the '**Factor(s):**' box by selecting **Add**.

(6) Select the transformed residuals (i.e., absresiduals), followed by **Add**, to insert the variable into the '**Dependent(s):**' box. Select **OK**. A graph of the data and the test output will appear within the '**Output Pane**' tab. A probability value greater than the usual criterion of $p > 0.05$ indicates that the data are homogeneous.

O.2.5 Weighting the Data

If Levene's test indicates that the residuals are heteroscedastic, and there is a significant change in variance across treatment levels (a distinct fan or 'V' shape, Figure O.2B), the data should be re-analyzed using weighted regression. The weight for a given treatment is the inverse of the variance of observations within that treatment. When performing the weighted regression, the standard error for the ICp (presented in SYSTAT as the asymptotic standard error ('A.S.E.'; see Figure O.3) is compared to that derived from the unweighted regression. If the difference is greater than 10%, then the weighted regression is selected as the best choice. However, if there is a significant change in variance across all treatment levels, and there is less than a 10% difference in the standard errors of the weighted and unweighted regressions[73], then the user should consult a statistician about alternative models, or the linear interpolation method could be used. The comparison between weighted and unweighted regression is completed for each of the selected models during the final selection of models and regression. Alternatively, if Levene's test indicates non-homogeneity, and the graphs of the residuals demonstrate a non-divergent pattern (e.g., Figure O.2C), an inappropriate model might have been selected. This would be another occasion to consult a statistician on alternate models.

[73] The value of 10% is only a rule of thumb based upon experience. Objective tests for the improvement due to weighting are available, but beyond the scope of this document. Weighting should be used only when necessary, since the procedure can introduce additional complications to the modelling procedure. If weighting is necessary, but the resulting estimates of parameters are nonsensical, a statistician should be consulted.

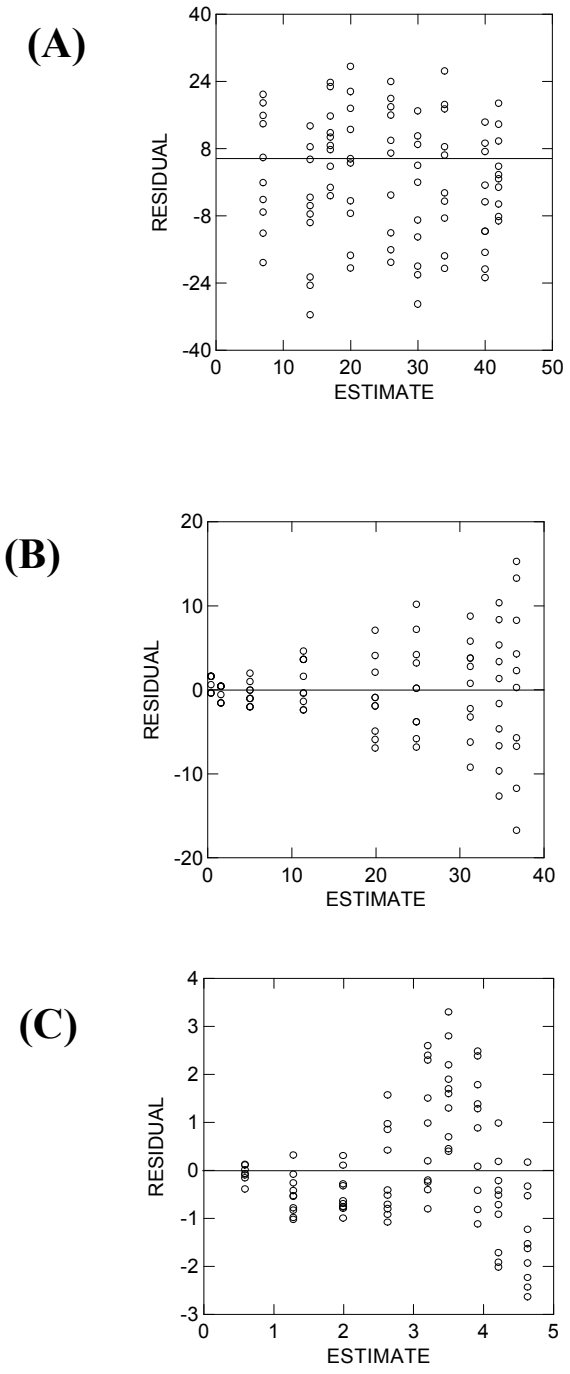


Figure O.2 **Residuals plotted against the predicted values.** (A) indicates homoscedasticity. Two types of heteroscedasticity are shown by (B) with a fan or 'V' shape which needs further examination using a weighted regression, and (C) which demonstrates a systematic lack of fit because an incorrect model was selected.

(1) Select **File**, **Open**, and then **Data**. Select the file containing the data-set to be weighted. Insert the two new variable names into the column heading by double-clicking on a variable name, which opens the '**Variable Properties**' window. In this window, insert an appropriate name for the variable of interest, select the variable type, and specify comments if desired. The two new column headings should indicate the variance for a particular variable (e.g., varjdrywt), and also the inverse of its variance (e.g., varinvsjdrywt). Save the data-file by selecting **File**, and then **Save**.

(2) Select **Data**, followed by **By Groups...**. Select the independent variable (logconc), followed by **Add**, to insert this variable into the '**Selected variable(s):**' box This will enable the determination of the desired variance by treatment level (i.e., "group"). Select **OK**.

(3) Select **Analysis**, **Descriptive Statistics**, and then **Basic Statistics...**. Select the variable of interest to be weighted (e.g., jdrywt), followed by **Add** to insert this into the '**Selected variable(s):**' box. Select **Variance** within the '**Options**' box, followed by **OK**. The desired variance, grouped by treatment level, will be displayed within the '**Output Pane**' tab of the main screen.

(4) Select **Data**, **By Groups...**, and then click on the box beside **Turn off**, and select **OK**, so that any following analysis will not be based on individual treatments, but on the entire set of data.

(5) Return to the data-file by selecting the '**Data Editor**' tab within the main screen. Transfer the variances for each concentration or treatment level to the corresponding concentration within the variance column (e.g., varjdrywt). Note that the variance is the same among replicates within a treatment.

(6) Select **Data**, **Transform**, and then **Let...**, and select the column heading containing the inverse of the variance (e.g., varinvsjdrywt) for the variable of interest, followed by **Variable** within the **'Add to**' box to insert the variable into the '**Variable:**' box. Select the '**Expression:**' box and type in '1/', and then select the column heading containing the variances (e.g., varjdrywt) of the variable of interest for each replicate and concentration, followed by **Expression** within the '**Add to**' box to insert the variable into the '**Expression:**' box. Select **OK**. The inverse of the variance for each replicate and concentration will be displayed in the appropriate column. Save the data by selecting **File**, and then **Save**.

(7) Select **File**, **Open**, and then **Command**. Open the file containing the command codes for estimating the parameters of the equation (Section O.2.3, step 2) for the same model selected for the *unweighted* analysis.

(8) Insert an additional row after the third line by typing 'weight=varinvsy', where 'y' is the dependent variable to be weighted (e.g., weight=varinvsjdrywt), as in the fourth line below.

```
nonlin
print=long
model drywts = t/(1+(0.25/0.75)*(logconc/x)^b)
weight=varinvsdrywts
save resid2/ resid
estimate/ start = 85, 0.6, 2 iter=200
use resid2
pplot residual
plot residual*logconc
plot residual*estimate
```

(9) Assign a new number for the residuals within the line entitled 'save resida' (where 'a' represents the assigned number).

```
SYSTAT Rectangular file C:\SYSTAT\STATAPP.SYS,
created Tue May 25, 2004 at 13:46:14, contains variables:
         CONC      REP      LOGCONC  JUVENILES  JDRYWT

 Iteration
 No.      Loss        G           X           B
   0 .452080D+04 .340000D+02 .400000D+00 .100000D+01
   1 .184579D+04 .328003D+02 .708478D+00 .157121D+01
   2 .157417D+04 .331384D+02 .696189D+00 .197718D+01
   3 .156445D+04 .329695D+02 .702780D+00 .211068D+01
   4 .156432D+04 .329461D+02 .703292D+00 .212794D+01
   5 .156432D+04 .329427D+02 .703387D+00 .212931D+01
   6 .156432D+04 .329424D+02 .703394D+00 .212941D+01

Dependent variable is JUVENILES

    Source   Sum-of-Squares    df  Mean-Square
 Regression       41208.683     3    13736.228
   Residual        1564.317    87       17.981

      Total       42773.000    90
Mean corrected    15140.456    89

       Raw  R-square (1-Residual/Total)        =        0.963
Mean corrected R-square (1-Residual/Corrected) =        0.897
         R(observed vs predicted) square       =        0.897

                                                  Wald Confidence Interval
Parameter         Estimate       A.S.E.    Param/ASE       Lower < 95%> Upper
 G                  32.942        1.031       31.952       30.893       34.992
 X                   0.703        0.031       22.898        0.642        0.764
 B                   2.129        0.229        9.299        1.674        2.585

          JUVENILES   JUVENILES
  Case     Observed    Predicted     Residual
     1       36.000       32.942        3.058
     2       31.000       32.942       -1.942
     3       22.000       32.942      -10.942
     4       25.000       32.942       -7.942
     5       39.000       32.942        6.058
     6       42.000       32.942        9.058
     .       ......       ......       ......
     .       ......       ......       ......
    86        2.000        0.337        1.663
    87        0.000        0.337       -0.337
    88        0.000        0.337       -0.337
    89        1.000        0.337        0.663
    90        0.000        0.337       -0.337

Asymptotic Correlation Matrix of Parameters

     G          X          B
G       1.000
X      -0.696      1.000
B      -0.611      0.566      1.000
```

residual mean square error

ICp, asymptotic standard error, and lower and upper 95% confidence limits

Figure O.3 Example of the initial output obtained with the Gompertz model in SYSTAT Version 11. The initial output provides the residual mean square error used to select the model of best choice, as well as the ICps, the standard error for the estimate, and the upper and lower 95% confidence limits. The number of cases displayed has been reduced for this diagram. The output within SYSTAT displays all cases, including the actual variable measurement and the corresponding predicted estimate and residual.

(10) Substitute the mean of the controls and the estimated ICp within the line entitled '*estimate/ start...*' (refer to Table O.2 for details on the substitution for each model). These estimates will be the same as those used for the unweighted analysis.

(11) Select **File**, and then **Submit Window** to run the commands. This will generate output of the iterations, the estimated parameters, and a list of the data-points with the corresponding predicted data-points and residuals, all within the '**Output Pane**' tab of the main screen. A preliminary graph of the estimated regression line will also be presented; this should be deleted. A normal probability plot and graphs of the residuals will also be presented.

(12) Proceed with the analysis as described in Section O.2.4 to ensure that all model assumptions have been met.

(13) Compare the weighted regression analysis with the unweighted regression analysis. Select the weighted regression if its standard error for the ICp is 10% less than that for the unweighted regression analysis.

O.2.6 The Presence of Outlier(s) and Unusual Observations

An outlier is a measurement that does not seem to fit the other values derived from the test. Outliers and unusual observations can be identified by examining the fit of the concentration-response curve to all data-points, and by examining the graphs of the residuals. If an outlier is observed, the general advice in Section 10.2 should be followed, which includes scrutinizing all experimental conditions and test records, whether hand-recorded or electronic, for human error. Identical examination must be given to all treatments, not just the one giving rise to the anomaly. The examination should also consider natural biological variation, and other biological reasons that might have caused the apparent anomaly. If an outlier is identified, analyses should be done with and without the outlier. Regardless of which analysis is regarded as definitive, a description of the data, outliers, and both analyses with their interpretive conclusions, must accompany the final report. If it seems that more than one outlier is present, the selected model should be re-assessed for appropriateness and alternatives considered.

The ANOVA function within SYSTAT can be performed as one method of determining whether or not the data contain outliers. However, ANOVA assumes that the residuals are normally distributed, and that assumption must be met before using the ANOVA. The presence of outliers can also be determined from the graphs of residuals, and by certain tests described in Section 10.2.

(1) Perform an Analysis of Variance (ANOVA) as described in Section O.4, to determine whether any outliers exist. They will be identified as a case number that corresponds with the row number in the SYSTAT data-file. The program uses the studentized residuals as an indication of outliers; values greater than three indicate the possibility of an outlier. This should be confirmed with the graphs of the residuals.

(2) If it is desired to perform an analysis without the anomalous datum, delete the value from the original data-table (file), and re-save the file under a *new* name (i.e., select **File**, and then **Save As...**). For example, the new file name might contain the letter 'o' (for outlier(s) removed) at the end of the file's original name.

(3) Repeat the regression analysis with the outlier(s) removed, using the same model and estimated parameters that were used with the outlier(s) present. An alternative model might be used for analysis if it resulted in a better fit and smaller residual mean square error. If the removal of the outlier(s) does not result in a significant change to both the residual mean square error and the ICp with its confidence intervals, then the investigator should use professional judgement on which analysis is superior. Justification for the choice must be provided, along with the records of alternative analyses.

O.2.7 Selection of the Most Appropriate Model

Once all of the contending models have been fit, each one should be assessed for normality, homogeneity of the residuals, and the residual mean square error. The model which meets all of the assumptions and has the smallest

residual mean square error (Figure O.3) should be selected as the most appropriate. If more than one model has the same residual mean square error, and all other factors are equivalent, the simplest model should be selected as the best choice. The residual mean square error is presented in the '**Output Pane**' tab just following the iterations, and preceding the estimates of parameters. If weighted and unweighted regressions were performed, the best one should be selected by the criterion provided in Section O.2.5. If none of the models provide a suitable fit to the data, the investigator should consult a statistician, or the data should be analyzed by the less desirable linear interpolation.

O.2.8 Creating the Concentration-Response Curve

Once an appropriate model has been selected, its concentration-response curve must be generated.

(1) Within the command editor window at the bottom of the screen, copy the model's equation from the command codes used to derive the estimates for the selected model. This is the equation after the '=' sign, in the third line of the command codes depicted in Table O.2. The equation should consist of the original alphabetic characters (e.g., t, b, h, etc.). The equation can be copied by highlighting the equation and selecting **Edit**, followed by **Copy** (or right-clicking the mouse and selecting **Copy**).

(2) Select **File**, **Open**, and then **Command** and open an existing graph command file (i.e., any file with '*.cmd') similar to the following example (or, if and as necessary, create a new one), using the logistic model. The first plot (i.e., 'plot') is a scatter plot of the dependent variable against the log concentration series. The second plot (i.e., 'fplot') is the regression equation, which is superimposed upon the scatter plot.

```
graph
begin
plot drywts*logconc/ title = 'Dry Mass of Barley Shoots', xlab = 'Log(mg boric acid/kg soil d.wt)',
ylab = 'Mass (mg)',
xmax = 2, xmin = 0, ymax = 90, ymin = 0
fplot y = 80.741/(1+(0.25/0.75)*(logconc/0.611)^2.533); xmin = 0,
xmax = 2, xlab = '' ymin = 0, ylab = '', ymax = 90
end
```

(3) Paste the previously copied equation in place of the pre-existing equation (as seen in the shaded area above) by highlighting the previous equation, and then selecting **Edit**, followed by **Paste** (or right-clicking the mouse and selecting **Paste**). Replace all of the alphabetical characters (e.g., t, b, h, x, a, etc.), together with the respective estimates, provided in the '**Output Pane**' tab generated by the application of the selected model.

(4) Type in the correct information within the line entitled 'plot y*logconc . . .', where 'y' is the dependent variable under study (e.g., drywts). Adjust the 'xmax' (i.e., the maximum log-concentration used) and 'ymax' (refer to Section O.2.1, Step 7) numerical values accordingly. Ensure that all 'xlab' and 'ylab' (i.e., axis labels) entries are correct, if not, then adjust accordingly. Ensure that all quotation marks and commas are placed within the command program as depicted in the previous example; SYSTAT is case- and space-insensitive.

'title' refers to the title of the graph
'xlab' refers to the x-axis label
'xmin' refers to the minimum value requested for the x-axis
'xmax' refers to the maximum value requested for the x-axis
'ylab' refers the y-axis label
'ymax' refers to the maximum value requested for the y-axis
'ymin' refers to the minimum value requested for the y-axis

The 'xmin', 'xmax', 'ymin', and 'ymax' must be the same for both plots to superimpose the regression line accurately on the scatter plot. An example of the final graph is provided in Figure O.1 for each of the five proposed models.

(5) Select **File**, then **Save As** to save the graph command codes in an appropriate working folder using the same coding used to generate the data-file, with an indication of the model to which the regression corresponds. Select **Save** to save the file.

(6) Select **File**, then **Submit Window** to process the command codes. A graph of the regression, using the parameters estimated for the selected model, will appear.

O.3 Determining Additional ICps

In some cases, it might be desirable to estimate a second ICp with another value for 'p'. Although the following section, and Figure O.1, are for determining an IC25, the models can be changed to suit any 'p' value (e.g., IC20).

(1) Select **File**, **Open**, then **Command**, and open the file corresponding to the command codes used to generate the estimates of parameters (refer to Table O.2 for the command codes for each model). Change the model equation such that it will calculate the desired ICp (e.g., IC25). Figure O.1 provides guidance on adjusting the models to calculate the IC25. Any ICp can be determined by modifying the fractions used in each model. For example, to calculate an IC20 using the logistic model, the equation for calculating an IC50, which is '*t/(1+(logconc/x)^b)*', would change to '*t/(1(0.20/0.80)*(logconc/x)^b)*' for calculating an IC20.

(2) Once the equation has been adjusted for the ICp of interest, follow each step outlined in Section O.2.3. However, substitute the initial estimate of ICp in the fifth line entitled '*estimate/ start=*' (refer to Figure O.1 for details on the substitution for each model). This is the value that was initially derived from an examination of the scatter plot or line graph. The model, once it converges, will provide a set of parameters from which the ICp and its 95% confidence limits, are reported (i.e., parameter 'x').

(3) Proceed with the analysis as described in Sections O.2.4 to O.2.8.

O.4 Analysis of Variance

(1) Select **File**, **Open**, and then **Data** to open the data-file containing all of the observations for the data-set.

(2) Select **Analysis**, **Analysis of Variance (ANOVA)**, and then **Estimate Model...**.

(3) Select the variable within which the data are to be grouped (e.g., logconc), and place this variable into the '**Factor(s):**' box by selecting **Add**.

(4) Select the variable of interest (e.g., jdrywt), followed by **Add**, to insert the variable into the '**Dependent(s):**' box.

(5) Select the box beside '**Save**' (bottom left-hand corner of the window called 'Analysis of Variance: Estimate Model') and scroll down the accompanying selections to choose **Residuals/Data**. Type an appropriate file name within the adjacent empty box to save the residuals (e.g., anova1). Select **OK**. A graph of the data and the generated output will appear within the '**Output Pane**' tab. At this point, any outlier(s), based on the studentized residuals, will also be identified (see Section O.2.6 for outliers).

(6) Assess the assumptions of normality and homogeneity of the residuals as in Section O.2.4, using the data-file that was created to save the Residuals/Data before conducting the ANOVA (i.e., anova1). Make the assessments using Shapiro-Wilk's and Levene's tests. The following coding can be used to examine the graphs of the residuals.

```
graph
use anova1
plot residual*logconc
plot residual*estimate
```

Hypothesis Testing

P.1 Statistical Methods

Hypothesis testing has been commonly used in the past for sublethal quantitative effects such as attained size. It is possible to transform some quantal data into a quantitative form suitable for analysis by hypothesis testing (Sections 2.92 and 2.9.3). Hypothesis testing can be applied directly to quantal data, without statistical difficulties, if numbers in a replicate are 100 or more, because the data become similar to quantitative distributions. For example, in the test with sea urchin eggs, the number of fertilized eggs is counted, among the first 100 or 200 encountered on a slide. The EC test method (1992f) acknowledges that this is a quantal effect, but with large enough numbers to be treated as a quantitative one. The procedure is not recommended for low numbers of observations per replicate, such as 40. The importance of large numbers is that the quantum jump in effect caused by one individual reacting within a group of 100, is only 1%, approaching a continuous distribution and satisfactory for quantitative techniques.

Statistical procedures for hypothesis testing are given in TOXSTAT (1996; WEST and Gulley, 1996), in CETIS (2001), and are explained with some guidance in USEPA (1994a), Newman (1995), and various EC sublethal test methods. The TOXSTAT and CETIS software are available commercially, and other suppliers provide broader general programs for computerized analysis. Procedures in the instructions that accompany the program should be followed. All providers of software packages modify the procedures to a greater or lesser extent in successive versions of the software.

A logarithmic scale is important in choosing the test concentrations; however, there is no need to ensure that logarithms of concentration are used in estimating NOEC/LOEC. The logarithms do not enter into the statistical analysis, because the statistical comparisons are among the observed *effects*. The groups could be identified just as well by using arbitrary numbers, letters, or names. In some cases the concentration is considered, for example, Williams' test considers order of concentration, although not the absolute magnitude.

P.2 Tests of Normality and Homogeneity of Variance

P.2.1 Shapiro-Wilk's Test for Normality

Calculations for this test are complicated and would be tedious if done by hand. TOXSTAT and other computer programs carry them out rapidly. The mathematical steps are shown in Newman (1995) and in an example by the USEPA (1995). The final step is comparison with a critical value (W) found in tables (Shapiro and Wilk, 1965; D'Agostino, 1986). The minimum sample size for this test is three.

An example of testing for normality can be based on the data shown in Table P.1. The data represent the weight gains in groups of late sac fry of rainbow trout, exposed to various concentrations of copper until they reached the early swim-up stage. Five concentrations and a control were tested. There were 12 fish per concentration, although 3 died in the highest concentration. These real data were obtained in the laboratory of Beak International, Inc. of Brampton, Ont.

In Table P.1, the two columns for “Weight gain” and “Residual” are relevant to the Shapiro-Wilk's test. Each value for a residual is simply the mean weight for the group, subtracted from the individual weight (see Glossary), and those residuals are the values that enter the Shapiro-Wilk analysis.

The calculations end with a critical value $W = 0.9836$, and the associated probability value is 0.5, i.e., very high. Compared to the usual criterion of $p > 0.05$, it is clear that the data are normally distributed. For a visual appreciation of such data, see Figure P.1.

Table P.1 Tabulation of toxicity data used as an example of assessing normality. The data represent weight gain of sac-fry of rainbow trout exposed to copper in water of 135 mg/L hardness. There are no replicates in this example, but in hypothesis testing , there would always be replicates. Data provided by Beak International, Inc.

Copper (µg/L)	Weight gain (mg)	Residual (mg)	Rank in group	Cumulative proportion	Probit
Control	66.7	-43.9	1	0.0769	3.5738
	101.5	-9.1	2	0.1538	3.9797
	102.7	-7.9	3	0.2308	4.2638
	103.7	-6.9	4	0.3077	4.4976
	105.0	-5.6	5	0.3846	4.7066
	109.3	-1.3	6	0.4615	4.9034
	111.7	1.1	7	0.5385	5.0967
	112.6	2.0	8	0.6154	5.2930
	122.2	11.6	9	0.6923	5.5018
	125.7	15.1	10	0.7692	5.7362
	128.9	18.3	11	0.8462	6.0203
	137.3	26.7	12	0.9231	6.4262
mean	110.6				
12	64.0	-25.4	1	0.0769	3.5738
	67.3	-22.1	2	0.1538	3.9797
	81.8	-7.6	3	0.2308	4.2638
	85.6	-3.8	4	0.3077	4.4976
	85.8	-3.6	5	0.3846	4.7066
	92.0	2.6	6	0.4615	4.9034
	92.0	2.6	7	0.5385	5.0967
	92.1	2.7	8	0.6154	5.2930
	96.5	7.1	9	0.6923	5.5018
	96.6	7.2	10	0.7692	5.7362
	105.4	16.0	11	0.8462	6.0203
	114.1	24.7	12	0.9231	6.4262
mean	89.4				
25	51.5	-41.3	1	0.0769	3.5738
	73.4	-19.4	2	0.1538	3.9797
	80.2	-12.6	3	0.2308	4.2638
	81.5	-11.3	4	0.3077	4.4976
	88.3	-4.5	5	0.3846	4.7066
	88.6	-4.2	6	0.4615	4.9034
	91.7	-1.1	7	0.5385	5.0967
	96.4	3.6	8	0.6154	5.2930
	109.0	16.2	9	0.6923	5.5018
	109.1	16.3	10	0.7692	5.7362
	112.6	19.8	11	0.8462	6.0203
	131.5	38.7	12	0.9231	6.4262
mean	92.8				
48	54.6	-28.1	1	0.0769	3.5738
	56.4	-26.3	2	0.1538	3.9797
	57.7	-25.0	3	0.2308	4.2638
	78.0	-4.7	4	0.3077	4.4976
	79.6	-3.1	5	0.3846	4.7066
	80.8	-1.9	6	0.4615	4.9034
	81.9	-0.8	7	0.5385	5.0967
	83.3	0.6	8	0.6154	5.2930
	97.4	14.8	9	0.6923	5.5018
	106.4	23.8	10	0.7692	5.7362
	107.8	25.1	11	0.8462	6.0203
	107.9	25.3	12	0.9231	6.4262
	82.7				
65	49.8	-16.1	1	0.0769	3.5738
	54.9	-10.9	2	0.1538	3.9797
	56.1	-9.7	3	0.2308	4.2638
	56.4	-9.4	4	0.3077	4.4976
	60.2	-3.6	5	0.3846	4.7066
	62.6	-3.2	6	0.4615	4.9034
	64.0	-1.8	7	0.5385	5.0967
	65.0	-0.8	8	0.6154	5.2930
	68.8	3.0	9	0.6923	5.5018
	69.2	3.4	10	0.7692	5.7362
	81.2	15.4	11	0.8462	6.0203
	102.0	36.2	12	0.9231	6.4262
	65.9				
91	11.7	-25.4	1	0.0769	3.5738
	13.5	-22.1	2	0.1538	3.9797
	19.1	-20.0	3	0.2308	4.2638
	41.3	-2.2	4	0.3077	4.4976
	45.6	6.5	5	0.3846	4.7066
	47.4	8.3	6	0.4615	4.9034
	56.6	17.5	7	0.5385	5.0967
	57.2	18.1	8	0.6154	5.2930
	59.2	20.1	9	0.6923	5.5018
	39.1				

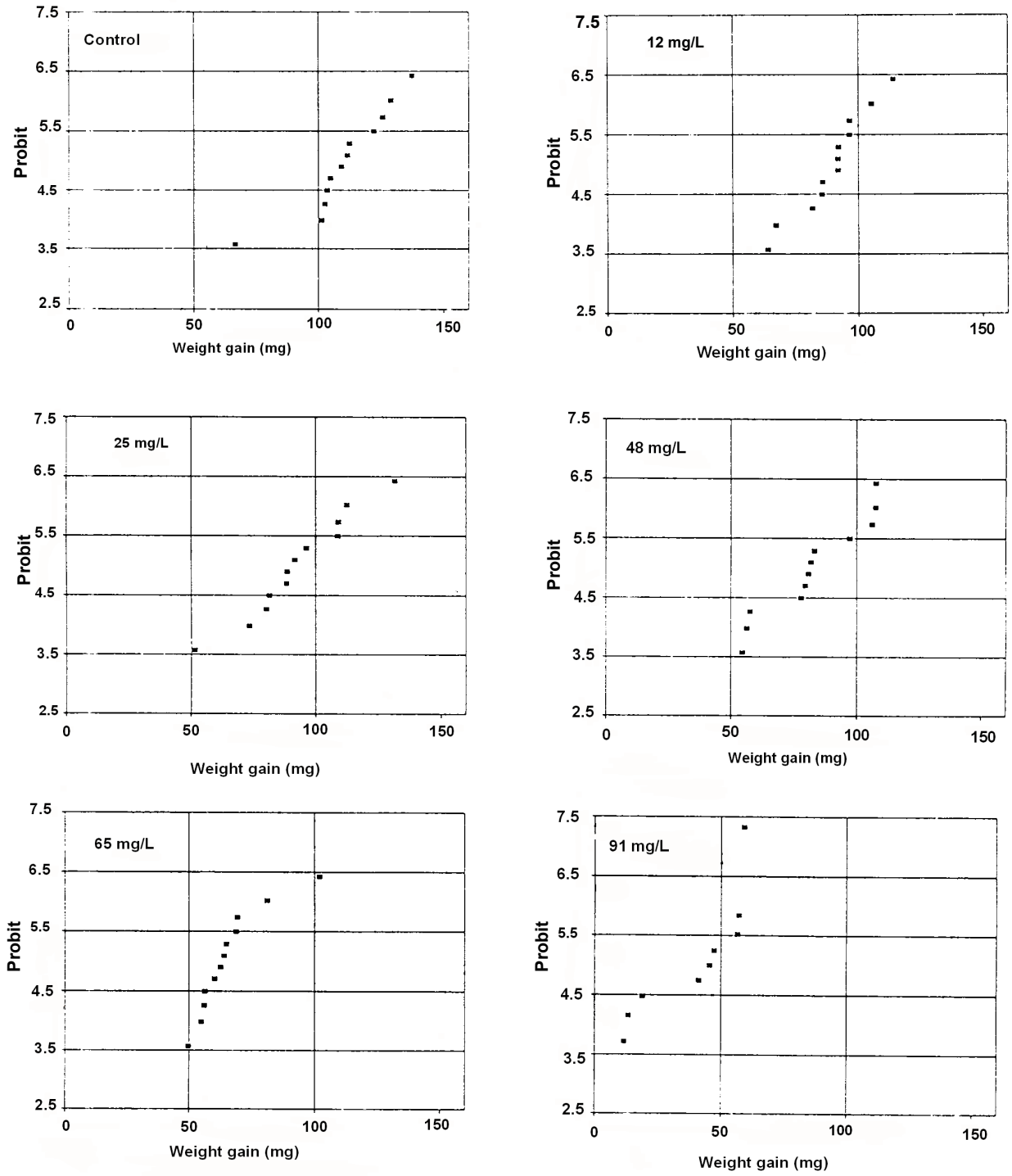


Figure P.1 **Plots to inspect apparent normality of distribution, for weight gains by rainbow trout sac-fry exposed to various copper concentrations.** Each panel represents the cumulative rank of each fry's weight gain within the distribution for 12 fry (on a vertical probability scale), plotted against the absolute weight gains (on an arithmetic scale). Three fry died in the highest concentration.

Investigators can assess the degree of non-conformity by the p-value offered in the computer program, or if necessary, in a table of critical values of "W" which should provide various probability levels from 0.01 upwards. Values of about 0.3 to 1.0 can be expected as the output (W) from the Shapiro-Wilk's test, with the lower value signifying considerable deviation from normality, and the value of 1.0 signifying little or no deviation.

Although normality tests could be conducted on the weights within each treatment, that is not recommended. The smaller sample sizes reduce the power of the test and increase the likelihood of a Type I error.

P.2.2 Plotting to Inspect Normality

The Shapiro-Wilk's test (Section P.2.1) is recommended for assessing normality, and should be the criterion for acceptance of the data. In addition, it could be instructive to plot graphs for visual appreciation of the distribution of data. Graphs should be based on the original data for a replicate or concentration. In cases of non-normality or non-homogeneity, the graph could reveal the apparent cause of failure to meet the requirements. It is not recommended that a graphic analysis by itself should be used to judge whether results are normal, because specialized graphic procedures are needed, as well as experience and skill for the subjective interpretation. For small sample sizes, there could be abrupt changes which could easily lead to over-interpretation. If visual assessment is done, the preferred methods in order are: quantile plots, box and whisker plots or stem and leaf plots, and frequency histograms.

Despite those qualifications, there is support in the literature for graphical appraisal of normality. Some support from OECD (2004) is described in Section 7.3.2 (footnote 54). Newman (1995) describes the procedure briefly and refers to detailed examples in Sokal and Rohlf (1981) and Miller (1986). Newman (1995) quotes Miller as writing "If a deviation from normality cannot be spotted by eye on probit paper, it is not worth worrying about." Beyond question; however, the eye that is used for spotting must be an experienced one.

Some examples of plotting can be given with the data from Table P.1 (Figure P.1). We already know that the data are normally distributed with a high probability value, from the Shapiro-Wilk's test in Section P.2.1, and so the panels of Figure P.1 illustrate relatively good data. It must be emphasized that *the test of normality is that the residuals are normally distributed.* Although in theory, if the effects are normally distributed at each concentration, the residuals should also be normally distributed, the actual tests of normality should be done on the residuals. Accordingly, Figure P.1 does not represent the visual assessment mentioned two paragraphs above (quantile plots, etc.) Figure P.1, to repeat, is merely a presentation of what relatively good data look like on probit plots.

The following outline is the procedure for calculating and plotting. The last three columns shown in Table P.1 are used. In other kinds of tests, "weight gain" would be replaced by whatever type of measurement was used.

- Within each concentration (or replicate if they exist), list the individual measurements in order from smallest to largest. (In this case, the measurements would be gain in weight for each of the 12 young fish.)

- Assign a number to each weight gain, indicating its rank in the list of twelve. For tied values, the average of the ranks should be used.

- For each weight gain, calculate the cumulative proportion of the data represented. Calculate these values by assuming that there is one additional value (12 + 1 = 13 for most of the treatments in Table P.1, and 9 + 1 for the highest concentration).

  Cumulative proportion = (Rank of the weight gain)/(number of measurements plus 1.0)

- Plot each of the cumulative proportions on a probit scale against its weight gain. (Alternatively, for each cumulative proportion, obtain the probit from a computer program or a table, and plot as the probit on an arithmetic scale, as in Table P.1 and Figure P.1.)

Weight gains of trout fry in Figure P.1, show a reasonably linear relationship in most cases, indicating probable normal distribution. There are some mild to moderate departures from normality, particularly for the individual showing least gain in the control, and the individuals showing the greatest gain (in 65 and 91 µg/L). Nevertheless, these data achieved a high degree of probability in the Shapiro-Wilk's test, and so Figure P.1 represents satisfactory normality of distribution.

If the experiment in Table P.1 were for hypothesis testing, there would be replicates. There would be an additional 12 sac fry in a separate test chamber for each replicate of a concentration. For plotting or testing the normality of residuals, each replicate would be plotted separately.

In some cases, a replicate observation would be a single number, such as the total weight or mean weight of all the individuals from a test chamber, which is the case for weight of larvae in the fathead minnow test. For tests of that design, the mean weight for a given replicate would be ranked among all the mean weights for the same concentration. The residuals from those rankings and mean weights would be plotted. If there were only two or three values, the plotting exercise would not be very revealing, and in fact, might be misleading. The Shapiro-Wilk's test would remain as the criterion.

P.2.3 Tests for Homogeneity of Variance

The method recommended here for assessing equality of variances is the test of Levene (1960), which is described in Snedecor and Cochrane (1980) but it is not presently included in software packages designed for environmental toxicology. The test of Bartlett (1937) is standard in the software packages but has a drawback (see following text). The test of O'Brien (1979) is somewhat superior to Levene's test, but is also absent from current statistical packages. Data based on proportions should not be put through these procedures.

All these tests determine whether the variances are equal in all the treatments, with a null hypothesis that there are no differences. If the variances differ substantively from treatment to treatment, the assumption of homogeneity required for a subsequent ANOVA is invalid. The tests of variance operate on the assumption that observations are normally distributed.

Bartlett's test is available in most software for environmental toxicology and is widely used. The test statistic is derived from the within-treatment variances and residual variances. The final comparison is with a critical value of chi-square, for the appropriate degrees of freedom and a selected probability value (α). For sample sizes less than five, a special table of critical values is used. Most investigators will allow a computer program to work through the calculations; the actual steps are shown in examples by Newman (1995) and USEPA (1995).

Bartlett's test is overly sensitive, if the data are not normally distributed and particularly if distributions are skewed. In such situations, a set of data might be erroneously rejected by the testing for homogeneity of variance.

Levene's test avoids that problem by using the average of the absolute deviations of an observation from its treatment mean, rather than the average of the squared deviations of within-treatment and residual variances. As mentioned, Levene's is not included as a standard test in software packages, nor is it mentioned or described in some textbooks (Zar, 1999; Newman, 1995). Levene's method could, however, be implemented by hand treatment of the data. Each observation should be recorded as an absolute deviation from the within-treatment mean. An ANOVA would then be performed on the recorded observations. The F-test for difference in the recorded observations would be a test of the assumption of homogeneity.

O'Brien's test has some superiority over Levene's test in certain technical mathematical aspects. However, it is even less easily available than Levene's, and is not explained in common texts (Snedecor and Cochran, 1980; Zar, 1999; Newman, 1995).

If the data being tested are proportions, then variances will differ with proportion and hence with treatment. Such quantal data should be analyzed by more appropriate methods than hypothesis testing (Section 4), or else suitably

transformed (Section 2.9.3). A warning about a particular difficulty with testing proportional data for homogeneity of variance was given by USEPA (1994d) [74], but the warning is not relevant if proportional effects are not included in hypothesis testing.

P.2.4 Robustness of Parametric Analysis and Decisions on its Use

If the data satisfactorily pass both the Shapiro-Wilk's and Levene's or Bartlett's tests, analysis should proceed with parametric methods, i.e., ANOVA.

If the data show inconsistencies and do not satisfy one or other of those tests, it might be possible to transform them statistically to meet the requirements for analysis. Transformation should be avoided, if possible, because there are complications and disadvantages as described in Section 2.9.2. If transformation is adopted, the set of modified data would be recycled through the tests for normality and homogeneity, to see if they now met the requirements. If so, analysis could proceed by standard parametric methods.

If the data cannot, even after transformation, satisfy both of those tests for the distribution of data, then analysis must be done using nonparametric methods (Figure 19). Computer packages usually assume that nonparametric analysis will be the only option, when one of the qualifying tests has failed.

However, a case can be made that ANOVA and the subsequent multiple-comparison tests are rather robust in the face of small deviations in normality and homogeneity. The tests for those characteristics function well for large samples, but might not do so for the small samples often found in environmental testing. The test for normality can be over-sensitive for unequal variances, and vice versa [75].

The relative robustness of ANOVA was described by Zar (1974) [76]. Newman (1995) cited work indicating that ANOVA produces realistic probabilities if the distribution of data is at least symmetrical and if the variances for the treatments are within three-fold of each other. One statistical program states that "An ANOVA can be valid even with departures from normality, especially when the number of replicates per group is large. If replicates are equal or nearly equal, heterogeneity of variance has little effect on the analysis" (TOXSTAT, 1996). Recent documents published by the USEPA also seem to have softened on this topic, with such wording as "If the tests fail ..., a nonparametric procedure ... may be more appropriate. However, the decision ... may be a judgement call, and a statistician should be consulted in selecting the analysis." (USEPA, 1995).
Accordingly, if the statistical tests for normality and homogeneity of variance indicate mild to moderate deviation from the requirements (i.e., marginal failure of a test), investigators might wish to consult a statistician about possible usefulness of parametric tests.

---

[74] If an investigator decided to directly analyze proportional (quantal) data by hypothesis testing, there is one situation which would have to be regulated by the investigator. It would result in unnecessary rejection of results for parametric testing, as the result of testing homogeneity of variance by Bartlett's or Levene's tests. An analysis of egg fertilization in sea urchins can be used as an example. It might happen that 100% fertilization occurred in each replicate of the control treatment. Similarly, there might possibly be 0% fertilization in each replicate of the highest concentration. In either of those situations, the variance for that treatment would be zero. When the test for homogeneity was run, the zero variance would result in rejection of the hypothesis of equal variances. If that situation occurred, the treatment with zero variance should *be omitted* from Bartlett's or Levene's test, and the ensuing estimate of the variance within treatments should be adopted (USEPA 1994d). If the other treatments met the condition of equal variances, then parametric analysis could proceed. In the subsequent analyses (ANOVA and multiple-comparison test), all treatments should be used including the one(s) with zero or 100% effects.

[75] Shapiro-Wilk's test for normality is sensitive to unequal variance, while Bartlett's, the test usually recommended for unequal variance, is known to be sensitive to non-normality. In view of this reciprocal sensitivity, investigators might have some justification if they did not adopt this pre-testing regime as an infallible article of faith.

[76] "Experience has shown that analyses of variance and *t* tests are usually robust enough to perform well even if the data deviate somewhat from the requirements of normality, homoscedasticity, and additivity. But severe deviations can lead to spurious conclusions." (Zar, 1974).

In this situation, some sublethal EC test methods recommend both parametric and nonparametric analysis, with the more sensitive (lower concentrations) of the two analyses providing the final estimates of toxicity [77]. That procedure is recommended here, and findings by both methods should be reported. Results of the Shapiro-Wilk's and O'Brien's (or Bartlett's) test should be submitted, along with a graph of raw results.

P.3 Analysis of Variance

For parametric testing, an *analysis of variance (ANOVA)* is carried out, with two main purposes. The initial purpose is to see if there is an overall difference between any two (or more) mean values for the various treatments (concentrations). This is done by testing the *null hypothesis (H_0)*, that there are no significant differences between mean values for the treatments. If a difference is found, the second purpose of the ANOVA is to obtain an estimate of the *error variance*; it will be used in further tests to see which particular concentrations differ.

The ANOVA makes use of (a) the total variance in the test, (b) the variance among concentrations, and (c) the variance within concentrations (i.e., among replicates). The estimates of variance are the "mean sum of squares" (fully expressed, the "mean squared deviations from the mean"), usually called the *mean square error*. These are obtained by dividing the sum of squares by the degrees of freedom. The sum of squares is obtained by subtracting each observation (replicate) from the mean of the category (concentration), squaring it, and adding all the squares together. The degrees of freedom are one less than the number of items in the category.

The analysis produces an output with the relevant values displayed in Table P.2. These hypothetical values would be for a test with five concentrations and three chambers (replicates) at each concentration[78].

Table P.2 The format of results from a hypothetical analysis of variance.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Total	2669	15 - 1 = 14	
Among concentrations	2046	5 - 1 = 4	511.5
Among chambers within concentrations	623	5 (3 - 1) = 10	62.3

For Table P.2, the real output of an ANOVA would probably label the three rows as "Total, Among, and Within" or "Total, Groups, and Error", rather than the explanatory labels shown. In the degrees of freedom column, only the 14, 4, and 10 would be shown, not the explanatory arithmetic. The values 623 and 10 in the third row could be obtained by subtraction.

If the mean square for "among concentrations" is larger than the mean square for "within concentrations", the null hypothesis might be untrue, i.e., significant difference in effect among two or more treatments. This is tested by

---

[77] The rationale for this is apparently based on the supposition that many parametric tests have greater power to detect effects, than do the corresponding nonparametric tests. They might detect a toxic effect for a set of data, even in the presence of minor irregularities, while a nonparametric analysis might fail to detect the effect.

[78] In some tests, effect might be measured for each of several organisms in a given chamber (replicate). The comparison of interest would be among the mean effects at different concentrations. This would be estimated by using the ratio of (a) the variation within the concentrations (i.e., among replicates), and (b) variation among concentrations. The measurements for individual organisms might be used in an ANOVA if, for some reason, it were desired to test for differences among replicates at the same concentration, as well as among concentrations. This would be a more complex "nested" ANOVA, described in statistics texts.

dividing “among” mean squares by “within” mean squares, the result being named “F”. If “F” exceeds a critical value, provided by the computer program, or found in tables, then there is a significant difference somewhere among treatments (concentrations).

In this hypothetical example, F = 511.5/62.3 = 8.2. The critical value of F for 4 and 10 degrees of freedom and p = 0.05 is 3.48. Since the calculated F is greater than the F found in the table, the null hypothesis is rejected and it is concluded that there is one or more difference(s) between concentrations.

The comparison of F with the critical value is only valid when the assumptions of the ANOVA have been met. This reflects back to the points raised in Section P.2.4.

If there is no significant difference shown by the ANOVA, the analysis is finished, the null hypothesis is accepted, and no toxicity has been demonstrated. If the null hypothesis is rejected, a difference exists and the statistical analysis proceeds to multiple-comparison tests (Section 7.5 and Section P.4), to decide which treatments differed from the control (and/or from which other treatments).

Generally these calculations are done by a computer program such as TOXSTAT, but it is possible to carry them out by hand using formulae shown in Newman (1995) or statistics texts (Zar, 1974; 1999).

One of the problems that might arise in ANOVA would be choosing the wrong value for the sum of squares of the “error”. If there were measurements on individual organisms within a replicate, and they had been entered into the analysis, Table P.2 would have additional numbers in another row at the bottom. Computer printouts would often label that row as “error”. An investigator might inadvertently use the mean square for that row in calculating F, which could be appropriate in some other experimental designs, as indicated in the preceding footnote, but not common. Usually the proper values can be identified fairly readily in the table printed out, and confirmation can be made by considering which line of the table has the correct number for degrees of freedom.

For testing hypotheses by ANOVA, it is highly desirable to have equal sample sizes (equal numbers of replicates per treatment). If there are inequalities, the analysis becomes more complex, but modern computer programs handle this easily and provide the correct error term for any subsequent multiple-comparison tests. Other important aspects of replication are given in Section 2.5. Interpretation and types of error are also relevant (Section 7.2.2).

P.4 Parametric Multi-comparison Tests

The use of multi-comparison tests is described in Section 7.5. Section P.4 provides some additional background information on the tests. Detailed guidance on multiple-comparison procedures is available (Hochberg and Tamhane, 1987).

P.4.1 Williams' Test

Williams’ test is a multiple-comparison test recommended for primary use in parametric analyses, after an analysis of variance has shown that a difference exists. It has a major advantage because, when comparing each treatment with the control, it takes into consideration the *order* of the groups according to increasing (or decreasing) concentration (Williams, 1972). Making use of this information increases the sensitivity of the test. Williams' test is provided in TOXCALC, TOXSTAT, and CETIS.

The superior sensitivity of Williams' test is demonstrated by an example. Crane and Godolphin (2000) compared precise test results from “laboratory 1” with variable results for “laboratory 2”. These were hypothetical observations on mortality, with three replicates, for a control and eight concentrations (as percent effluent, 1.0, 2.2, 4.6, 10, 22, 46, 60, and 100). Data were transformed as square roots and analyzed by ANOVA and several multiple-concentration tests.

The differences were striking. Not only were the calculated NOECs surprisingly different for the two laboratories, but also for the different statistical tests (Table P.3). Williams' test was the most sensitive of the four tests by

Table P.3 Differences in NOECs calculated by various multiple-comparison tests. The NOECs represent percent effluent, for hypothetical data from laboratory 1 with precise data, and laboratory 2 with variable data, as presented by Crane and Godolphin (2000).

Multiple-comparison test	NOECs for Lab. 1	NOECs for Lab. 2
Williams'	1.0	2.2
Dunnett's	2.2	22
Bonferroni t-test	2.2	22
Tukey's	10	46

factors of 2 to 20 times. It was particularly effective in establishing a low concentration for the variable data of laboratory 2.

Williams' test proceeds in a stepwise manner. It starts by comparing the effect of the first-ranked sample (e.g., highest concentration) with the control effect, then that of the next ordered sample, until no difference is found. By that process, the test identifies the lowest concentration associated with a significant mean effect in a test group.

Williams' test is related to the t-test and shares the same assumptions. Effects must be approximately normally distributed, variances within concentrations must be equal, and observations must be independent. Those requirements should have been met for the preceding ANOVA. If the requirements are not met, the nonparametric stream would be appropriate, using Shirley's test (Section P.5.3) as the alternative for Williams' test.

The test must operate on a monotonic series, i.e., each successive mean effect is either (a) equal to or smaller than the previous one, or else (b) equal to or larger than the previous one. In case of irregularities, there is a smoothing procedure, which might have to be applied by hand. The correction assigns the same mean effect to the two aberrant mean effects in the series. The correction can be made more than once if necessary, but in the usual series for a toxicity test, such equalization of groups could lose an important part of the test's ability to discriminate. These situations will be seen easily when an investigator inspects or plots the original data; if it exists the investigator should apply Williams' test and also another multiple-comparison test, to check for anomalous results.

Williams' test will function for equal or unequal numbers of observations contributing to the mean value of the control and each treatment. Normally the calculated error term is obtained with a computer program. If a particular computer package cannot deal with unequal numbers of observations among the treatments, the adjustments can be made by hand. There is a choice of simple formula for balanced or unbalanced data (Williams, 1972).

The critical value for a given set of data, corresponding to the error degrees of freedom, can be obtained from tables in Williams (1971; 1972). For unbalanced data, the critical values should be obtained from the tabulation of Hochberg and Tamhane (1987). Comparing the calculated test statistic to the critical value, the first one that is less than the critical value is significantly different from the control.

P.4.2 Dunnett's Test

Dunnett's test is a standard test which compares the mean effect at each treatment with the mean effect in the control. Dunnett's test is given prominence in TOXSTAT and most current methods from the USA.[79] However, Williams' test is recommended here instead, for Environment Canada tests that have ordered data (e.g., successive concentrations). Dunnett's is less powerful than Williams' for establishing LOEC, because it ignores any order in

[79] The computer software for Dunnett's test is available at http://www.epa.gov/nerleerd/stat2.htm.

the data (Table P.3). Also, it controls for the experiment-wise error rate rather than the pairwise error, when comparing any treatment with the control.

Dunnett's test is, however, the appropriate choice for making a comparison with the control when there is no intrinsic ordering of the treatments, i.e., no gradient is expected. This could occur, for example, in sediment-testing, if there were materials from a number of different places, all tested in replicate, but only at one concentration (i.e., full-strength).

Dunnett's test requires a normal distribution of data; it represents an extension of the t-test (Dunnett, 1955; 1964). Dunnett's is usually set up in computer software packages to carry out a *one-tailed test* of significance, which fits the expected situation that the measurements in the test concentrations will all be in the same direction from the measurement in the control. Dunnett's test gives conservative results (tendency not to identify differences) for the normal one-tailed tests.

Dunnett's test is usually applied to experiments which have an equal number of observations at each treatment, and the older available software packages offer only that option. Sometimes unequal numbers could occur, such as more observations in the control. The best remedy would be to download a recent version of the "modified" Dunnett's test (see footnote 79). There is also a suitable modification explained in Newman (1995), and worked examples are found in USEPA (1995). The other options for unequal numbers of observations are the Dunn-Sidak test or the Bonferroni-adjusted t-test.

P.4.3 Dunn-Sidak and Bonferroni Adjustments for Unequal Replicates

The modified Dunnett's test is recommended for comparing each treatment with the control, when numbers of observations are unequal. If the adaptation of that test for unequal observations was not available, the Dunn-Sidak test could be used. The Bonferroni adjustment is mentioned because it is used in the United States, but it has no particular advantages and need not be considered for use.

Both the Dunn-Sidak and Bonferroni adaptations compare the mean of each treatment with the mean for the control. Neither is very powerful compared to Williams' test, i.e., real differences might not be distinguished. The Bonferroni adaptation is currently standard in software packages, the Dunn-Sidak is provided in CETIS, TOXCALC, and TOXSTAT but might not be available in some packages. An example of the Bonferroni adaptation is worked in USEPA (1995).

The Dunn-Sidak and Bonferroni adaptations are based on the t-test, with an adjustment of the critical values of *t*, to correct for a multiple comparison. Repeated pairwise comparisons with a normal t-test could result in a Type I error (Section 7.2.2). The required adjustments are made automatically in computer packages, and over-compensate somewhat. The table of critical values for the Dunn-Sidak test can be inspected, if desired, in Newman (1995).

P.4.4 Pairwise Comparison Tests

There are tests for checking the difference between all possible pairs of treatments. Although this is not likely to be needed for most toxicity tests, it could be of interest for field testing or comparing various locations. *Fisher's Least Significant Difference (LSD)* is akin to the t-test, and is recommended. It has the favourable feature of controlling pairwise, rather than experiment-wise, Type I error. The LSD can be used for equal or unequal replication. It is intended for only a few of all possible comparisons in a set of data, comparisons which would be specified in advance, and in that respect is similar to other multiple-comparison tests. The LSD is included in the computer package SYSTAT and some others that can be used in toxicity work, and is described in some textbooks (Steel and Torrie, 1980; Steel *et al.*, 1997). Orientation on use of the LSD is provided in Section D.2.2 of USEPA and USACE (1994).

As substitutes for LSD, *Tukey's test* and the *Student-Newman-Keuls test* (SNK) are commonly available in software packages for environmental toxicology. Tukey's test can cope with unequal sample sizes although equality is desirable. The sensitivity of Tukey's is low (Table P.3).

P.5 Nonparametric Methods for Estimating NOEC

If the data from a test cannot meet the requirements for normality and/or homogeneity of variance, even with transformation, they should be analyzed by nonparametric methods, using the tests described here and in Section 7.5.2. These nonparametric options are strong tools for data that are not normally distributed. However, in general they would be less powerful in detecting a toxic effect than corresponding parametric tests, if they were used on normally distributed data.

Certain nonparametric methods require at least four replicates and sometimes five [80]. This is acknowledged in the methods documents for specific sublethal tests published by Environment Canada.

P.5.1 Initial Tests of Hypothesis

Many of the nonparametric multi-comparison tests are self-sufficient and do not have an absolute need to be preceded by a test that would be analogous to ANOVA. Omitting that initial hypothesis testing step has been common practice in toxicological work. However, this document recommends that many of the nonparametric multi-comparison tests should be preceded by hypothesis-testing (see Figure 4). In these cases, the analysis should proceed to multiple comparison, only if the initial test rejects the hypothesis of no difference among treatments. The reason for this is to avoid Type I errors in the multiple comparison. In other words, the aim is to avoid declaring a significant difference between two treatments when the difference is the result of chance, an event that is expected in one out of twenty comparisons for the usual p-value of 0.05). In statistical parlance, the multiple-comparison test is being *protected* by the initial screening test of hypothesis. This two-stage testing is a conservative approach, and conceivably, it might occasionally result in failing to detect a difference that is real (Type II error).

Descriptions of three of these tests follow, for use with different types of nonparametric data (Figure 4). They are the nonparametric equivalents of an ANOVA (Zar, 1999) and indicate whether or not there is at least one difference among the treatment effects. These tests do not indicate which one is different from which others. Their particular use in different situations is shown in Figure 4 and will be indicated in following sections.

The Kruskal-Wallis Rank Sum test (hereafter called the Kruskal-Wallis test) was described by Kruskal and Wallis (1952). It is sometimes provided in software packages (TOXSTAT, 1996) as if it were only a multiple-comparison test, the nonparametric equivalent of Tukey's test. However, this test can be used for hypothesis testing (ANOVA analogue), and also as a multiple-comparison test.

The Fligner-Wolfe test is a rank sum test that can be used to test a null hypothesis of no effect (Fligner and Wolfe, 1982). It tests the null hypothesis that none of the treatment medians differ from the control median, against the alternative hypothesis that all treatment medians are greater than the control median. This alternative hypothesis is different from the usual one with such tests, and is quite explicit. One serious effect of this is that the test is not appropriate when some treatments (concentrations) result in a higher measured effect and some result in a lower measured effect. Therefore the test is not suitable for toxicity tests displaying hormesis, in which cases, the Kruskal-Wallis test should be used. The other limitation of the Fligner-Wolfe test is easily overcome. If the treatments in a toxicity test result in lower values for the effect measured, all those values should be multiplied by minus unity (-1).

The Jonckheere-Terpstra test (Jonckheere, 1954) also performs as a nonparametric analogue of ANOVA, and has very good power. The null hypothesis is that all medians are equal and the alternative hypothesis is a little different

---

[80] The requirement for four replicates could be a problem. A test might have been designed with three replicates, primarily for calculating a point estimate as recommended in this document. If the investigator wished to calculate the NOEC/LOEC, that could also be done with parametric methods. If, however, the results deviated from normality, and required analysis by nonparametric methods, then the investigator might not be able to determine the NOEC and LOEC, depending on the particular nonparametric test. Recent test methods published by Environment Canada require four replicates for hypothesis testing, but that would not be enough for use of Shirley's test.

from usual -- that the treatments are ordered. Accordingly, it is very suitable for toxicity tests. Although available in some major statistical software packages, this test unfortunately, is not yet included in software for toxicology, and the hand calculations are very tedious and time-consuming.

P.5.2 A General Multi-Comparison Test

The Edwards-Berry test (Edwards and Berry, 1987) is a multiple-comparison test which could follow any of the three previously mentioned procedures for hypothesis testing. If the null hypothesis were rejected by whichever test was being used, then the Edwards-Berry test would be suitable for any of the situations described in the sections that follow. Regrettably, this test is not readily available in software packages yet, but could be used as it becomes more widely available. The Edwards-Berry test uses boot-strapping to develop an empirical distribution for the data. Because of that approach, the method can handle most configurations of data, whether balanced or not. A critical value is produced which protects the family-wise comparison error rate.

P.5.3 Ordered Data -- Shirley's Test or Pairwise Comparison

Shirley's test is a very desirable nonparametric method. It parallels the parametric Williams' test, and considers the ranking of the concentrations, in their increasing (or decreasing) order. It is used to compare effects with the control, and is not preceded by a hypothesis-testing procedure (i.e., no nonparametric analogue of ANOVA is used, see Figure 7.1). It is adaptable for unequal replication. Shirley's is an extension of the Kruskal-Wallis test (Section P.5.1), but can be expected to yield results like Williams' test. The test assumes that the effects are monotonically decreasing, and if not, they are smoothed as in Williams' test. The within-treatment sample size must be five or more.

Shirley's test ranks the groups for degree of effect by using the mean values of effects in the control and treatment groups. The actual mean values are not used in the analysis as they would be in Williams' test. The control effect(s) is/are ranked in the same series as the treatments (test concentrations). The test compares the mean rank for a given concentration, with the mean rank of the control. The variance is the nonparametric variance of the ranked observations. The procedure works on a rank sum basis. The rank of the highest concentration is compared with that of the control. If that is significant, the comparison proceeds to the next lower concentration until no difference is found.

Shirley's test should be used when it becomes available, but unfortunately it is not found in most computer packages for toxicology, nor in some general statistical packages such as SPSS (1996; 2001). Nor is the method described in some standard textbooks. The test can be carried out by hand, although it is tedious. If the test is not available, an investigator requiring a nonparametric test could use a pairwise comparison of the ordered data (Section P.5.3) if the appropriate tests were available. The other possibility for comparison with the control only, would be to use the options for a non-ordered set of data, starting with the Fligner-Wolfe test (Section P.5.4).

Pairwise comparison of ordered data starts with hypothesis testing, using the Jonckheere-Terpstra test (Section P.5.1). If the null hypothesis of no difference was rejected, analysis would proceed with the ***Hayter-Stone test*** (Hayter and Stone, 1991). This multiple-range test can deal with equal or unequal replication. Tables of critical values are available for large or small samples, if there is equal replication (i.e., balanced data). For unbalanced data, there is a more limited availability of the critical values. At the time of writing, the tables of critical values have been provided only for smaller sets of unbalanced data, including three treatments or fewer, and with no more than seven replicates.

The computer software is not readily available, for either the Jonckheere-Terpstra or Hayter-Stone tests.

P.5.4 Comparing Non-ordered Data with the Control

In such a non-ordered situation, the *Fligner-Wolfe test* (Section P.5.1) is recommended to test the null hypothesis of no differences from the control. If that is not available in suitable computer software, the Kruskal-Wallis test could be used. If the null hypothesis is rejected, and the data are balanced, the recommended first choice for a multiple-comparison with the control is the ***Nemenyi-Damico-Wolfe test*** (Damico and Wolfe, 1987).

A second choice for the multiple-comparison test is the ***Wilcoxon Rank Sum test*** which is generally available, and handles unequal replication. The Wilcoxon test arises from procedures and critical values developed by a number of statisticians (Newman, 1995).

The Wilcoxon test functions in a similar fashion to Steel's Many-One Rank test (see following text). For a given concentration, the differences between test measurements and their corresponding controls are ranked. A plus or minus sign is given to each ranking, according to the nature of the difference from the control. Positive ranks are summed, and negative ranks also. The smaller of the positive and negative sum is compared with known critical values to determine whether a significant difference exists between test effect and control effect. Repeating this for each concentration yields an estimate of NOEC and LOEC. This test is generally available in computer software programs. An example is worked in USEPA (1995).

A third choice is ***Steel's Many-One Rank test*** (Steel, 1959; 1961) which is offered in most statistical packages and is called by several names. An example of the test is worked in USEPA (1995). As available in software programs, this test is suitable only for data with an equal number of observations in each treatment and the control(s). At least four observations (replicates) are required. Computer packages provide a one-tailed test, that is, all the samples with toxicant are presumed to cause effects the same as, or greater than the control. Being the nonparametric equivalent of Dunnett's test, Steel's test can be used in comparisons such as sediment testing, as mentioned previously.

The method has ranking at its core. A set of replicate measurements for a given concentration (say, four mean weights from four replicates) is listed together with the four measurements from the control. The eight mean measurements are ordered by rank (increasing magnitude). The rankings of the test measurements are added together and the rankings of the control measurements are also added. The lower of the two sums of ranks is compared with a critical value from a standard table of critical values. The test measurements for this concentration are declared either different or not different from control measurements. This procedure of listing along with control values, is repeated for each test concentration. At the end, the investigator knows which concentrations have an effect that is significantly different from the control (further details in Newman, 1995). There is a modification for the case in which all test concentrations have the same number of observations, but the control has a different number. Although this modification is not available in the usual software packages for environmental tests, it is described in Newman (1995).

P.5.5 Pairwise Comparison of Non-ordered Data

The first choice for a multiple-comparison test is the ***Critchlow-Fligner-Steel-Dwass test***, commonly called the *Critchlow-Fligner test* (Critchlow and Fligner, 1991).[81] This test could be used if the preceding Kruskal-Wallis test rejected the hypothesis that all treatments showed median effects that were equal.

The test compares the results from each treatment with those from each other treatment including the control, and it indicates whether the medians are equal or different. The Critchlow-Fligner test would be preceded by the Kruskal-Wallis test (Section P.5.1), and would only be used if that test rejected the null hypothesis. The Critchlow-Fligner test is suitable for equal or unequal replication among the treatments. This is a two-sided test for difference, i.e., a difference could be one treatment showing effects greater in magnitude than another treatment, or lesser in magnitude. A given comparison of two treatments is not influenced by effects measured in other treatments; this is a very desirable feature in a nonparametric multiple-comparison test (Miller, 1981). The test controls the experiment-wise error rate, and there is a low probability that one will declare a difference in two treatments, when there is none.

---

[81] Steel and Dwass proposed such a pairwise test independently, but it was only for balanced data. The publication by Critchlow and Fligner (1991) extended the test to cover unbalanced results, and so all four names are suitably associated with the test method.

The Critchlow-Fligner test is not included in the usual packages of computer software, and would have to be taken from its description in Critchlow and Fligner (1991). Tables of critical values are only available for a limited number of sample sizes, although the additional required tables could be generated.

Steel's Pairwise test (Steel, 1960) is a second choice for a multiple-comparison test, and it is suitable for balanced data. If the set of data was unbalanced, and the Critchlow-Fligner test was not available, the Kruskal-Wallis test would be pressed into double duty. First it would be used to test the null hypothesis and in the case of rejection, the same test would be used for multiple comparisons, to find which treatment effect(s) differed from which others.

Statistical Differences Among EC50s

Section 9.5.2 provided Equation 9 as a potential method for carrying out a chi-square test for differences among three or more EC50s.

$$\chi^2 = \sum_{i=1}^{k} w_i \left(\log(EC50_i) - \left(\frac{\sum_{i=1}^{k} w_i \log(EC50_i)}{\sum_{i=1}^{k} w_i} \right) \right)^2 \qquad [Equation\ \ 9]$$

Equation 9 may be revised as in Equation Q.1 to make it appear less formidable, and to more easily illustrate the steps in the calculations.

$$\chi^2 = \sum_{i=1}^{k} w_i \left(\log(EC50_i) - c \right)^2 \qquad [Equation\ \ Q.1]$$

In turn, *c* may be defined as in Equation Q.2, and b may be calculated as shown.

$$c = \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} w_i} \qquad [Equation\ \ Q.2]$$

$$b = w \times \log(EC50) \qquad w = \left(\frac{1}{SE_{(logEC50)}} \right)^2$$

Example calculations are shown in Table Q.1. The comparison is based on three of the EC50s which were previously used as examples in footnote 63 of Section 9.5.1. The EC50s and confidence limits are 8 (5.3, 12), 11 (7.3, 16.3), and 15 (10, 22.5), used here as examples A, B, and C, respectively. The EC50s are changed back to logarithms for calculations in Table Q.1. Standard Errors (SE) are also calculated on a logarithmic scale and shown in the second row of the table. In the rows below that, the calculations are carried out according to the formulae above.

There is at least one significant difference among the three EC50s, but it is not known where the difference(s) lie(s). A suitable multiple-comparison test has not yet been defined, but might be developed (Zajdlik, in prep.). Still, the procedure in Table Q.1 could already be useful. It will at least define situations in which a significant difference does *not* exist, possibly avoiding needless speculation on the cause of differences which were, in fact, not significant.

Table Q.1 Example calculations to calculate chi-square for testing significant differences among three EC50s.

	Toxicity tests			Sigma
	A	B	C	(= sum)
log EC50	0.90309	1.041399	1.17609	
SE of logEC50	0.06392	0.06284	0.06343	
w = (1/SE) squared	244.7224	253.2706	248.5211	746.5141
b = w * logEC50	221.0064	263.7541	292.2835	777.0440
c = Sigma b / Sigma w				1.0409
logEC50 - c	-0.13781	0.00050	0.13519	
previous item squared	0.01899	0.00000	0.01828	
w * previous item (= contributions to chi-square)	4.64744	0.00006	4.54237	9.190
Critical value of chi-square for 3-1 deg. of freedom, and p 0.05 = 5.991				
Calculated value (9.190) is higher than critical value, therefore there is at least one significant difference among the three EC50s.				

Another difficulty in using this procedure at present, is that an investigator is seldom informed of the value of SE, by the usual statistical programs for estimating EC50. Confidence limits of the EC50 are, however, provided, and statistics textbooks describe a general relationship between the confidence interval and the standard error (e.g., Zar, 1999). That relationship is for conventional types of means and limits, and it is uncertain whether it can be used for the toxicity data. This matter might be resolved by Zajdlik (in prep.)

Median and Quartiles

The *median* and *quartiles* are useful and apparently simple mathematical concepts. However, it can become confusing when attempting to apply quartiles to laboratory data. This appendix explains the confusion but might not remedy it.

The median is uniformly defined in the literature as the "middle" number (item) in a series of numbers that is ranked from lowest to highest numerical value (or alternatively, from highest to lowest). With an odd number of items in the series, the median is the numerical value of the middle item/number. With an even number of items, the median is the average of the numerical values of the two middle items. In either case, the median fulfils its purpose of dividing the series so that *half of the items in the ranked series precede the median and half of them follow it*.

Quartiles have a similar purpose of further equal division of a ranked series of numbers. One rule of thumb, the *interquartile range*, i.e., the numerical difference between the *first quartile* and the *third quartile* can be used to identify possible outliers. As described in the glossary, one-quarter of a ranked series of numbers would occur before the first quartile, and three-quarters would occur before the third quartile.

It is recommended here, that in picking quartiles for a series, an investigator should choose the most reasonable values for satisfying the general definitions above, i.e., the one-quarter and three-quarters definitions. Sometimes this becomes difficult to decide in short series, and in such cases the concept of quartiles becomes less useful and might best be avoided.

The dilemma in short series is increased because various mathematical authorities specify different systems for identifying or calculating the quartiles. As many as eight variations in method have been listed (web site www.xycoon.com/quartiles). Outlines of the usual methods follow, but contradictions can be seen by searching apparently academic web sites on the internet for "quartile" and "statistics".

The most commonly encountered version of the quartiles will probably be the one that is used in recent versions of the spreadsheet programs EXCEL and QUATTRO-PRO. The lower quartile is derived from the formula $L = 1/4 (n + 3)$, and the upper quartile from $U = 1/4 (3n + 1)$. (The symbol "n" is the number of items in the series.) If the result from either formula is a whole number (integer), then that number indicates which item in the list is the quartile. (For example, if the answer is 3, pick the third item in the list of numbers). If the result from a formula includes decimals, then it tells the item in the list and the proportion that must be calculated between that item and the next item in the list. (For example, if the answer is 3.75, pick the third item in the list and interpolate three-quarters of the way between the third item and the fourth item.) EXCEL describes the latter situation: "If a quartile falls between two discrete values in the list, a fractional value is determined by linear interpolation". It might be noted that if there is an odd number of items in the list, this method includes the median value in the lower half of the list when determining the lower quartile, and also in the upper half when determining the upper quartile. Including the median in that way was the procedure used by the statistician John Tukey when he was defining quartiles so that they could be calculated by simple methods.

This spreadsheet method can be applied to the very short list which was used as an example in the footnote of Section 10.2 (the series 20, 24, 28, 34, 40). Applying the formula for L, the answer is 2.0, so the lower quartile is the second item in the list, i.e., 24. For U, the answer from the formula is 4.0, so the upper quartile is the fourth item, i.e., 34. Another example would be a list of eight items: 4, 5, 7, 9, 10, 12, 13, 16. By the formula, $L = 2.75$, so the lower quartile is 6.5. U equals 6.25, so the upper quartile is 12.25. The decimals are retained as part of the values of the quartiles.

Statistics Canada recommends picking the quartiles in the same manner used for the median, and provides formulae (web site www.statcan.ca). The first quartile is the middle item of the ranked observations *below* the median. The third quartile is the middle item of ranked observations *above* the median. (The median is not included in either series.) For an even number of items, the quartile is calculated by the same method used for the median. Their example is a 12-item series: 1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57. The median is the average of the 6th and 7th items, (24 + 28)/2 = 26. Below that, there are six items with values from 1 to 24. The first quartile is the average of the third and fourth values, 15 and 19 = 17. Similarly, the third quartile is 42. Statscan gives no advice on procedure if the quartile is a decimal value, but presumably that value would be accepted as for the median.

Some authoritative sources indicate that the quartiles must be actual values from the ranked series of numbers. That simple procedure was followed in the example in the footnote 64; Section 10.2. (Quartiles of 24 and 34 were selected for the series 20, 24, 28, 34, 40.) Use of actual values from the series is an end-result from formulae described by Zar (1974), which are similar to the spreadsheet formulae shown previously. The expression (n + 1)/4 identifies which item is the first quartile. Similarly, 3(n + 1)/4, or 0.75(n + 1) identifies the item in the series which is the third quartile. If the result is not an integer (e.g the 1.5th item), the next higher integer is selected (e.g., the 2nd item in the series). That selection can bring problems in a short series like the one shown previously (20, 24, 28, 34, 40); the first quartile is the second item in the series (24), but the third quartile would be the fifth item in the series, i.e., 40. The fifth item is the last number in the series and scarcely fulfils the purpose of having one-quarter of the values following it.

A variant of the preceding method was once used in some early versions of the Microsoft spreadsheet program EXCEL. If the calculation resulted in a decimal value of × .51 or higher, the next higher item in the series was selected (e.g., for 1.51, the second item). If the calculated result was × .49 or lower, the preceding item was selected (e.g., for 1.49, the first item). If × .50 was obtained, that value was used as the quartile (e.g., 1.5). If the result was an integer, that item was used (e.g., 2.0, use the second item).

Sometimes advice is found, to omit the "+ 1" from those formulae described by Zar (1974) and given above. The first quartile would be described by n/4 and the third quartile by 3n/4, moving up to the next higher integer if required. For the five-item series used as an example above, that would give quartiles of the second and fourth items, i.e., 24 and 34 as used in the footnote 64; Section 10.2.

Some authorities provide greater exactness in the selection process by allowing values that could potentially have occurred in the series. For example in a series of integers, a quartile would be an integer, but might not have actually occurred in the series.

In the face of this differing advice, the recommendation here is that an investigator who must pick first and third quartiles, should pick the most sensible values that satisfy the definition in the glossary, i.e., to divide the series of numbers into four equal parts. For long series of numbers, the conflicting advice will not create a problem in identifying quartiles; all methods will give similar answers, and the spreadsheet method will be satisfactory. For very short series, the use of quartiles is not recommended.